Fusing Pre-trained Language Models with Multimodal Prompts through Reinforcement Learning

Youngjae Yu🟩* Jiwan Chung🟩* Heeseung Yun🟨 Jack Hessel🟥
Jae Sung Park🟥🟦 Ximing Lu🟥🟦 Rowan Zellers🟪
Prithviraj Ammanabrolu🟥 Ronan Le Bras🟥 Gunhee Kim🟨 Yejin Choi🟥🟦
🟥 Allen Institute for Artificial Intelligence 🟪 OpenAI
🟩 Department of Artificial Intelligence, Yonsei University
🟨 Department of Computer Science and Engineering, Seoul National University
🟦 Paul G. Allen School of Computer Science, University of Washington

Abstract

*Language models are capable of commonsense reasoning: while domain-specific models can learn from explicit knowledge (*e.g. *commonsense graphs [6], ethical norms [25]), and larger models like GPT-3 [7] manifest broad commonsense reasoning capacity. Can their knowledge be extended to* multimodal *inputs such as images and audio without paired domain data? In this work, we propose* ESPER *(Extending Sensory PErception with Reinforcement learning) which enables text-only pretrained models to address multimodal tasks such as visual commonsense reasoning. Our key novelty is to use reinforcement learning to align multimodal inputs to language model generations without direct supervision: for example, our reward optimization relies only on cosine similarity derived from CLIP [52] and requires no additional paired (image, text) data. Experiments demonstrate that* ESPER *outperforms baselines and prior work on a variety of multimodal text generation tasks ranging from captioning to commonsense reasoning; these include a new benchmark we collect and release, the ESP dataset, which tasks models with generating the text of several different domains for each image. Our code and data are publicly released at* https://github.com/JiwanChung/esper.

1. Introduction

Collecting multimodal training data for new domains can be a Herculean task. Not only is it costly to assemble multimodal data, but also curated datasets cannot completely cover a broad range of skills, knowledge, and form (*e.g.* free text, triplets, graphs, etc.). Ideally, we want to endow multimodal models with diverse reasoning capacity (*e.g.* ethics [25], commonsense [57], etc.) without undertaking a separate multimodal annotation effort each time.

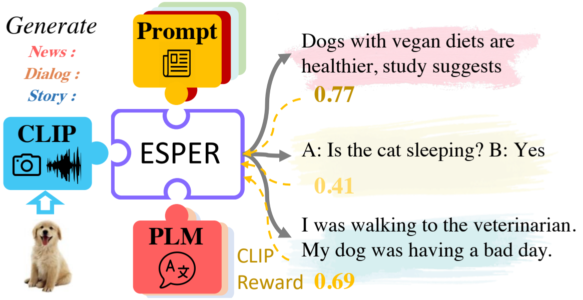

Generate
News :
Dialog :
Story :
Prompt
CLIP
ESPER
PLM
Dogs with vegan diets are healthier, study suggests
0.77
A: Is the cat sleeping? B: Yes
0.41
I was walking to the veterinarian. My dog was having a bad day.
CLIP
Reward 0.69

Figure 1. The intuition of ESPER, Extending Sensory PErception with Reinforcement learning. To better align knowledge in CLIP and pretrained language models (PLM), we use CLIP as a reward for the pairs of images and self-generated text.

In this work, we propose Extending Sensory PErception with Reinforcement learning(ESPER), a new framework that enables a pre-trained language model to accept multimodal inputs like images and audio. ESPER extends diverse skills embodied by the pre-trained language model to similarly diverse multimodal capabilities, all without requiring additional visually paired data. In a zero-shot fashion, our model generates text conditioned on an image: using this interface, we show that ESPER is capable of a diverse range of skills, including visual commonsense [50], news [39], dialogues [60], blog-style posts [28], and stories [22].

ESPER combines insights from two previously disjoint lines of work: *multimodal prompt tuning* and *reinforcement learning*. Like prior multimodal prompt tuning work, ESPER starts from a base language-only model (e.g., GPT-2 [53], COMET [6]), keeps most of its parameters frozen,

[*] denotes equal contribution

and trains a small number of encoder parameters to map visual features into the embedding space of the language model [40, 45, 68]. *Unlike* prior works, however, ESPER does not train these parameters using maximum likelihood estimation over a dataset of aligned (image, caption) pairs. Instead, it uses a reinforcement learning objective. During training, the model is first queried for completions conditioned on visual features. Then, the lightweight vision-to-text encoder is updated using proximal policy optimization (PPO) [59] to maximize a similarity score computed with a secondary pre-trained image-caption model, CLIP [52]. As a result, the frozen language model can interpret the multimodal inputs in the same context as the text embedding space without additional human-annotated paired data.

Reinforcement learning has two advantages over maximum likelihood objectives. First, RL bypasses the costly process of collecting multimodal paired data by using a reward model. This relaxation is especially favorable when image/audio groundings rarely exist (*e.g*. ethics [25] and knowledge graph [57]). The second major advantage of RL is the maintenance of generalizability. Similar to prior work, we freeze the parameters of the language model, which helps keep its reasoning capacities. However, ESPER goes a step further: Tsimpoukelli et al. [68] and Mokady et al. [45] fine-tune their lightweight adapters using paired visual-linguistic datasets such as Conceptual Captions [61] or COCO Captions [38]. Because these literal scene descriptions cannot match the textual variety of the large-scale corpus GPT-2 is trained on, the supervised models may generate less richly styled language or be capable of as diverse reasoning over input contexts [33, 73].

We experimentally compare ESPER to two classes of prior methods that seek to adapt language models to accept visual inputs: (1) maximum likelihood prompt tuning [45, 68]; and (2) decoding-time methods [67] that post-process token probabilities of a frozen language model according to estimated image similarity. For zero-shot image/audio captioning, we find that ESPER outperforms all prior unsupervised methods, both in terms of text quality (e.g., 14.6 point improvement in CIDEr over Laina et al. [34] in COCO unpaired captioning) and inference speed (e.g., 100× speedup vs. Tewel et al. [67], which relies on per-token gradient optimization over partial decodings).

In addition, we show that ESPER can efficiently adapt without paired resources on visual commonsense reasoning [6, 50], visual news [39], visual dialogue [11], and our new zero-shot multimodal generation benchmark named ESP (Evaluation for Styled Prompt dataset), which tests the model's capability to generate text of different domains for the *same* image. Furthermore, ESPER also shows the capability to learn about audio inputs using an audio reward. We hope the strong performance of ESPER presented here will encourage researchers to consider RL-based training for future multimodal prompt tuning work.

2. Method

ESPER consists of three components: 1) CLIP's non-generative image/text encoders [52];[1] 2) a left-to-right language generator such as GPT-2 [53] or COMET [6]; and 3) an encoder that projects multimodal inputs into the subword embedding space of the language generator.[2] During training, CLIP and the language generator's parameters are frozen; gradients are back-propagated through the frozen language model to train the encoder parameters. We employ reinforcement learning (specifically, PPO [59]) to derive these gradients: our reward is the similarity of the sampled text to the input image, as estimated by CLIP. After RL training, we evaluate ESPER in various zero-shot scenarios. We use CLIP ViT-B/32 and GPT-2-base (12-layer) as defaults for our experiments. In this setting, ESPER features 8M trainable and 300M untrainable parameters.

2.1. Architecture

CLIP. Radford et al. [52]'s Contrastive Language Image Pre-trained (CLIP) encoder plays two roles in our framework: first, as a feature extractor for the input images, and second, as an alignment reward between the images and the model-generated text. First, the fixed CLIP image encoder $CLIP\text{-}I$ extracts single vector feature from the image x^i; second, the fixed CLIP text encoder $CLIP\text{-}T$ is applied to text samples the model generates; finally, we use the cosine similarity of these two vectors as the reward signal.

Encoder. The encoder F_ϕ is the only trainable module in ESPER. Given the vector representation of an image x^i extracted using CLIP, the module outputs a series of vectors of length k to be passed on to the language model, i.e.,

$$h^i = h^i_1, \ldots, h^i_k = F_\phi(CLIP\text{-}I(x^i)) \tag{1}$$

The output image representations h^i work as the multimodal prompt and are concatenated to the embedded word representations. For fair comparison in later experiments, we use the same multimodal encoder architecture as CLIP-Cap [45]: a lightweight, two-layer Multi-Layer Perceptron (MLP), and set $k = 10$.

Pre-trained Language Model. ESPER employs a pre-trained deep autoregressive language model such as GPT-2 [53] as the backbone. Autoregressive language models

[1] While we describe image modeling here, we also experiment with audio/text encoders, specifically Wav2CLIP [74], in § 3.4 that extend ESPER to audio inputs.

[2] In principle, any model architecture with the same APIs could be used, e.g., ALIGN [24] could be substituted for CLIP, or T5 [54] could be substituted for GPT-2

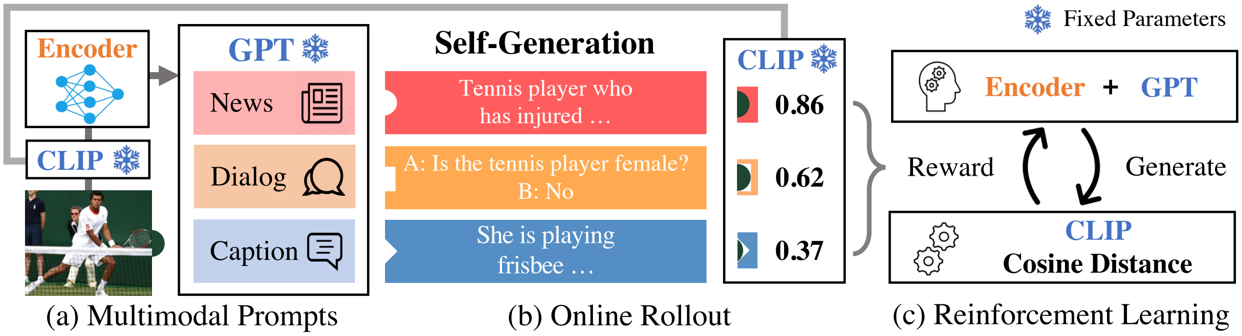


Figure 2. Illustration of the proposed model, ESPER. We use a pre-trained language model (*e.g.* GPT-2 [53]) as the language generator.

parameterize likelihood of a text sequence y factored as text tokens y_j with length l using autoregressive decomposition.

$$p_\theta(y) = \prod_{j=1}^{l} p_\theta(y_j | y_{j'<j}) \tag{2}$$

Inspired by prompt tuning in the text-only domain [40], we concatenate h^i with the output of GPT-2's text embedding lookup layer to prefix the conditioned text generation:

$$p_\theta(y^i | h^i) = \prod_{j=1}^{l} p_\theta(y^i_j | h^i, y^i_{j'<j}) \tag{3}$$

The initial text prompt can be as short as a single subword token for free-form training or contain task-specific templates for zero-shot adaption to downstream tasks. The parameters of the language model θ are kept frozen.

2.2. Training

Reinforcement Learning. We propose to view CLIP as a black-box model and apply reinforcement learning to maximize cosine similarity between the input image and generated text as a reward.[3] From the RL perspective, our language generator can be viewed as a policy, which produces actions (in the form of generations) given states (in the form of text+image prompts). We use the clipped version of Proximal Policy Optimization (PPO-clip) [59, 63] as the RL algorithm to optimize the reward: specifically, we adapt the implementation of PPO from Stiennon et al. [63]. Stiennon et al. include an additional reward term that penalizes the KL divergence between the RL policy and the original policy to ensure the generation stays fluent and meaningful. Our value model has the same architecture as the policy model (see Sec. 2.1). We use random sampling with temperature 0.7 during training.

Given an input image x and the corresponding generated text y, our reward is given by:

$$\alpha \left(\frac{CLIP\text{-}I(x)}{||CLIP\text{-}I(x)||} \cdot \frac{CLIP\text{-}T(y)}{||CLIP\text{-}T(y)||} \right) + \beta \tag{4}$$

where $\alpha = 50, \beta = -10$ are fixed normalizing factors that empirically cause the reward function to have roughly zero mean and unit variance during training. We defer the detail of the PPO-clip algorithm to Appendix C.

Language Model Stability. Reward hacking can potentially occur [30] if the agent discovers incoherent texts that nonetheless achieve high rewards. To prevent this, we incorporate auxiliary rewards to stabilize the training process. First, we compute the KL divergence between p_θ and a separate (fixed) text-only GPT-2 model to maintain language generation capability. We also find it beneficial to consider text-only likelihood as an additional reward. Finally, as reported in previous literature [20, 72], language models often falsely assign high likelihoods to repetitive phrases. We introduce an explicit repetition penalty against this phenomenon. Refer to Appendix C for the details.

2.3. Adaptation on Pretrained Language Model

ESPER can also adapt to domain-specific language models. We use the unpaired text data to train the domain-specific backbones such as COMET [6] via standard supervised learning. Then, we build the multimodal text generator (*e.g.* ESPER-COMET) by finetuning the backbone using RL *without any paired data*. ESPER-Domain denotes an overarching term for the domain-specific types of ESPER. In all our experiments, we build the text backbones by finetuning GPT-2 [4] with domain-specific text data such as ATOMIC [57], News [39] and Dialog [39], alongside the corresponding prompts (e.g., `"news:"`, `"dialog:"`, etc.). On the other hand, ESPER also extends the zero-shot

[3] Because CLIP does not provide differentiable per-token feedback (as the model is only differentiable given a full caption) it's not possible to do gradient-based updates directly; we quantitatively compare against Tewel et al. [67] who apply CLIP to partial decodings.

[4] Again, ESPER can utilize any generative text model architecture, e.g., T5 [54].

SocialMedia : @janew Look how shiny is my daughter's hair! The AAA conditioner really works!

News : It's been hard on kids since the lockdowns during the pandemic. They're so used to going to school, it's hard for them to stay home all the time. …

Blog : Teaching my daughter to brush her hair by herself has been a challenge. Now, I have finally gotten to the point where I sit and talk with her while she combs her hair instead of me doing it for her.

Instruction : Brush hair in even strokes. Use a detangling spray if there are knots that are hard to get out. Spraying only a few spurts of the detangling spray will be sufficient.

Story : Abbie realized she had to look good. She took out her comb and start strengthening her hair. It was easy for she usually does it without help. The hair she strengthened flattened smoothly.

Figure 3. A sample in Evaluation for Styled Prompt dataset (ESP dataset).

Model	Domain	B@4	M	C	Time
Pseudo-Align [34]	✓	5.2	15.5	29.4	-
RSA [21]	✓	7.6	13.5	31.8	-
Unpaired [34]	✓	19.3	20.1	63.6	-
ZeroCap [67]		2.6	11.5	14.6	65s
ZeroCap-CaptionLM	✓	7.0	15.4	34.5	65s
CLIPRe [64]	✓	4.9	11.4	13.6	-
MAGIC [64]	✓	12.9	17.4	49.3	3s
ESPER-GPT		6.3	13.3	29.1	0.65s
ESPER-CaptionLM	✓	**21.9**	**21.9**	**78.2**	0.65s

Table 1. Unpaired captioning experiments in COCO test split. B@4 denotes Bleu-4, M METEOR, and C CIDEr score. Running time entails each step of inference, including image loading and feature extraction. *Domain* indicates domain-specific text-only pre-training. CLIPRe is a retrieval-based approach using CLIP.

Model	Zero-shot	B@4	M	C
CLIPCap-MLP		27.4	22.4	94.4
CLIPCap-Full		32.2	27.1	108.4
ESPER-CaptionLM	✓	21.9	21.9	78.2
ESPER$_{\text{Init}}$-MLP		31.2	25.4	103.1
ESPER$_{\text{Init}}$-Full		**33.1**	**27.7**	**111.1**

Table 2. Finetuning experiment in COCO Captions test split. We omit the zero-shot baselines here for readability.

adaptability of general language models (GPT-2) to multimodal inputs. ESPER-GPT is an instance of the general-purpose models that do not utilize task-specific text data.

3. Experiments

Our main goal is to extend the diverse knowledge in pre-trained language models to multimodal domain. Experiments in Sec. 3.1 test whether ESPER can align vision and language, and those in Sec. 3.2 and Sec. 3.3 check that ESPER maintains textual diversity in the backbone. As GPT is a general-purpose language model, we want to show that ESPER can likewise work as a general-purpose multimodal text generator on diverse tasks such as commonsense reasoning, news, and dialogue. Finally, we extend ESPER to another modality of audio in Sec. 3.4.

3.1. Evaluation of Visual Alignment

We first evaluate strength of the alignment between an input image and the generated text in ESPER using the MSCOCO captioning corpus [38]. While ESPER could benefit from a more diverse set of unpaired images, for fair comparisons with the baselines, we limit our data to COCO training set images (*unpaired* with their captions).

3.1.1 Zero-Shot Image Captioning

Following previous works on unpaired captioning [13, 34], we split the pairing between image and caption and train them separately using ESPER for unsupervised evaluation. We split COCO Captions dataset [38] with Karpathy split [26]. Models are evaluated with BLEU-4 [49], METEOR [2], and CIDEr [69]. The models in Table 1 use greedy decoding to generate descriptions at inference time.

In Table 1, without any explicitly paired MSCOCO data, we show that ESPER outperforms a variety of prior works in unpaired captioning [21, 34], and CLIP-based decoding methods [64, 67]. Domain-specific language generators improve conditional generations: ESPER-GPT, which does not know COCO caption text, falls behind ESPER-CaptionLM (which is pre-trained on unaligned COCO captions, with the prefix `caption:`). Also, at inference time, ESPER is faster than decoding time methods like Tewel et al. [67].

3.1.2 RL helps, even in supervised finetuning

As our encoder shares the same architecture with MLP-variant CLIPCap [45], we can directly evaluate the benefit of ESPER's RL training, even if supervised data is available. We experiment two supervised variants: ESPER$_{\text{Init}}$-MLP, which finetunes only the encoder, and ESPER$_{\text{Init}}$-Full, which finetunes the encoder and language model jointly. Table 2 shows that initializing with ESPER's RL-trained encoder outperforms random initialization when performing usual maximum likelihood training; this promising result shows that RL and MLE training can complement each other.

(a) Visual Commonsense Graph (VCG)				
Model	Zero-shot	B@4	M	C
Retrieval [52]	✓	0.3	7.0	5.6
ZeroCap-COMET [67]	✓	3.0	10.0	13.1
CLIPCap [45]	✓	0.0	6.4	0.9
VisualCOMET [50]		12.5	10.7	16.5
Text-VCGLM	✓	9.9	9.5	12.8
ESPER-VCGLM	✓	**13.0**	**10.5**	**16.4**

(b) COCO Captions + COMET						
	Val			Test		
Model	B@4	M	C	B@4	M	C
Retrieval [52]	15.2	19.5	17.3	15.3	19.5	18.0
ZeroCap-COMET [67]	8.8	13.1	8.0	8.4	13.2	8.0
CLIPCap [45]	10.9	12.3	17.3	10.5	12.4	17.6
Text-COMET	17.8	18.9	3.3	17.7	18.9	3.3
ESPER-COMET	**28.3**	**23.6**	**28.9**	**28.3**	**23.6**	**29.7**

Table 3. Commonsense reasoning experiments in (a) unpaired Visual Commonsense Graph validation split [50] and (b) COCO Captions with commonsense extension of text-only COMET [6].

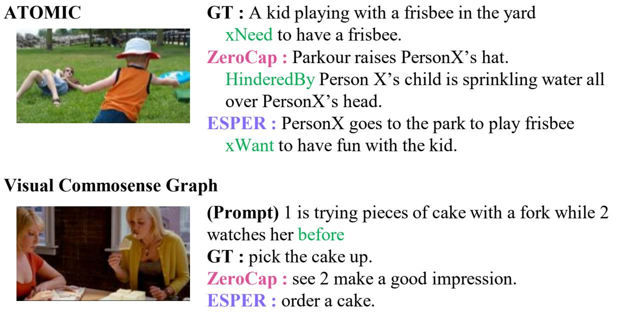
(a) News				
Model	Zero-shot	B@4	M	C
Show Attend Tell [75]		0.7	4.1	12.2
Text-Only	✓	0.2	2.7	1.3
ZeroCap-News [67]	✓	0.3	0.2	0.0
CLIPCap [45]	✓	0.2	3.8	1.6
ESPER-NewsLM	✓	0.8	4.4	4.6
$\text{ESPER}_{\text{Init}}$-MLP		**1.3**	**4.8**	**15.7**

(b) Dialog				
Model	Zero-shot	NDCG	MRR	R@1
ViLBERT [42]	✓	11.6	6.9	2.6
ViLBERT-Head		19.7	9.8	3.4
Text-Only	✓	19.3	18.3	5.7
ESPER-DialogLM	✓	**22.3**	**25.7**	**14.6**

Table 4. Downstream task evaluation in (a) VisualNews [39] test split and (b) VisDial [11] validation split.[5] NDCG denotes Normalized Discounted Cumulative Gain, MRR Mean Reciprocal Rank and R@1 Recall at top 1.

Model	B@4	M	C
Retrieval [52]	0.71	6.46	2.49
ZeroCap-CaptionLM [67]	1.29	5.91	6.21
ESPER-GPT-2-Audio	0.32	4.62	2.84
ESPER-CaptionLM-Audio	**3.40**	**9.47**	**7.92**

Table 5. Audio alignment experiment in AudioCaps [27] test split.



GT : A kid playing with a frisbee in the yard
xNeed to have a frisbee.
ZeroCap : Parkour raises PersonX's hat.
HinderedBy Person X's child is sprinkling water all over PersonX's head.
ESPER : PersonX goes to the park to play frisbee
xWant to have fun with the kid.

(Prompt) 1 is trying pieces of cake with a fork while 2 watches her before
GT : pick the cake up.
ZeroCap : see 2 make a good impression.
ESPER : order a cake.

Figure 4. Examples for zero-shot commonsense reasoning experiments; ATOMIC [57] and Visual Commonsense Graph [50].

3.2. Fusing Domain-specific Language Models

Going beyond standard image captioning setups, we evaluate ESPER's capacity to adapt to text generators with domain-specific knowledge such as commonsense. First, we extend i) commonsense graphs to multimodal inputs. We then evaluate ii) news and iii) dialogue domains, which have existing public corpora.

3.2.1 Visual Commonsense Graph

While general language models (*e.g.* GPT-2) embody implicit commonsense, commonsense knowledge graphs offer explicit structures to represent commonsense. For instance, ATOMIC [57] connects two everyday events with nine types of If-then relations (*e.g.* cause and effects). We evaluate ESPER on commonsense graphs with a text-only dataset (ATOMIC) and a visual-language dataset (Visual Commonsense Graph [50]). Figure 4 shows examples and model output of ESPER in the selected commonsense datasets.

We first use Visual Commonsense Graph (VCG) dataset [50] to show that ESPER can extend commonsense graphs to visual inputs. Given an image and the corresponding event description, VCG evaluate commonsense reasoning capability to generate text description on what happens before, what will happen after the event, and why the event took place. As in other experiments, ESPER-VCG uses unpaired data with image and caption decoupled. We compare ESPER-VCG against a supervised baseline of VisualCOMET (trained with VCG) as well as a text-only finetuned GPT-2. Note that all compared methods use event description as the only text context, discarding place annotation.

Results in Table 3-(a) show that ESPER improves over the text-only baseline and even performs on par with VisualCOMET, a baseline trained with image-caption pair information not provided in ESPER-VCG. Hence, ESPER training can substitute supervised training in commonsense graphs.

Next, we turn to a more complex task of adapting a text-only dataset without any visual annotation. Here, we fuse commonsense knowledge in ATOMIC [57] to visual

[5] We use the validation set for evaluation for fair comparison against previous zero-shot baseline results [47]

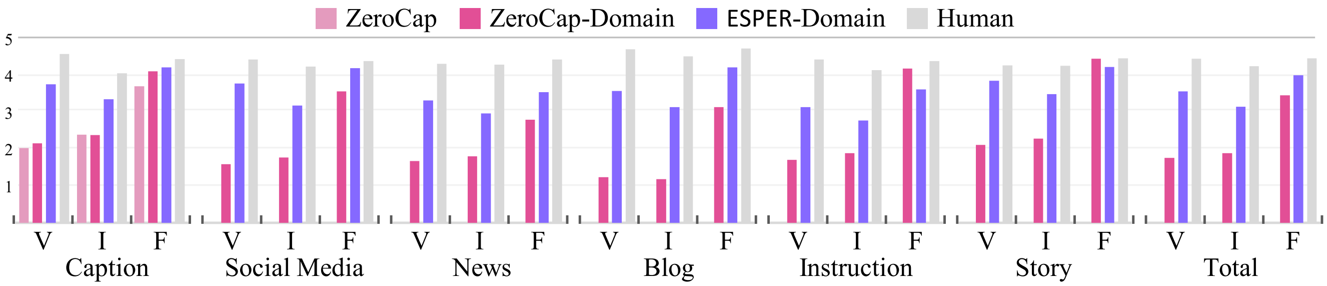


Figure 5. Human evaluation of captions for each domain prompt. We take the average of 5-point Likert-scale ratings from three annotators. V denotes visual relevance, I is informativeness, and F is fluency. Domain denotes domain-specific backbones described in Section 2.3.

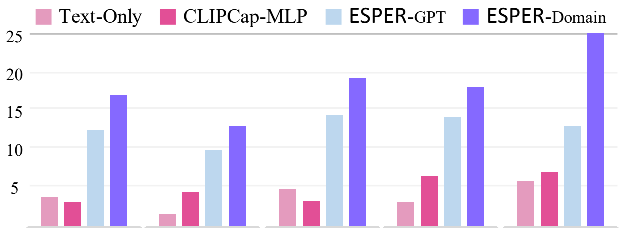


Figure 6. ESP dataset experiment. We report CIDEr in this plot.

stimuli. For evaluation purposes, we instead build visually aligned validation set using ATOMIC and COCO captions. We defer the detail of evaluation set in appendix.

In Table 3-(b), we show that ESPER outperforms retrieval-based and CLIP-based decoding methods. Thus, ESPER trains a visually-conditioned commonsense model successfully even when the text data is collected outside of a visual context.

3.2.2 Visual News

VisualNews [39] includes 1.08 million news images along with the associated image captions and articles. For a fair comparison, we compare against models that rely only on image inputs, [6] e.g., Show Attend Tell [75], from Liu et al. [39]. We also include the text-only backbone as another baseline (Text-Only).

Results are in Table 4-(a): zero-shot ESPER outperforms not only the text-only baseline but also the supervised baseline in Bleu-4 and METEOR scores. However, it lags behind the supervised model in CIDEr. While various proper nouns appear in news articles, CLIP has no knowledge of most of them. Since CIDEr takes rarity of terms into account, this difference in data extends to performance degradation in news texts generated from ESPER. By finetuning the adaptor, ESPER overcomes the knowledge gap and exceeds the baselines even in CIDEr.

3.2.3 Visual Dialogue

VisDial [11] is a dataset of iterative dialogues conditioned on an image. The model output should be the ranking of the given 100 next-response candidates. The reported metrics (NDCG, MRR, and R@1) compare the model ordering of the next response with the human ordering: refer to the dataset paper [11] for details on the evaluation scheme. After training ESPER with the unpaired dialogue-domain generator, we rank the answer candidates by their likelihood. The baselines consist of zero-shot ViLBERT [42] and frozen ViLBERT finetuned with a linear head.

Table 4-(b) shows the VisDial dataset re-ranking results. Zero-shot ESPER improves the baselines by a large margin. It even outperforms the supervised ViLBERT-Head, showing that ESPER can discern likely visual dialogues.

3.3. From One Image to Many Domains

While we observe that ESPER-GPT can generate diverse image-related texts, we still need to prove that this textual diversity is controllable by text prompts; a null hypothesis is that there are identifiable and consistent features found, e.g., only in news images, and that ESPER cannot produce diverse captions for *the same* image.

ESP dataset. To benchmark ESPER-GPT's capability to generate diverse domain-specific language from the *same* image, we collect and release a novel dataset, ESP dataset (Evaluation for Styled Prompt dataset): a benchmark for zero-shot domain-conditional caption generation. It comprises 4.8k captions from 1k images in the COCO Captions test set [38]. We collect five text domains with everyday usage: blog, social media, instruction, story, and news, as illustrated in Figure 3. We defer the dataset details and the collection process to Appendix D and E, respectively.

Automatic evaluations on ESP dataset Figure 6 shows that ESPER generates conditional text depending on the domain prompts. ESPER outperforms CLIPCap-MLP [45],

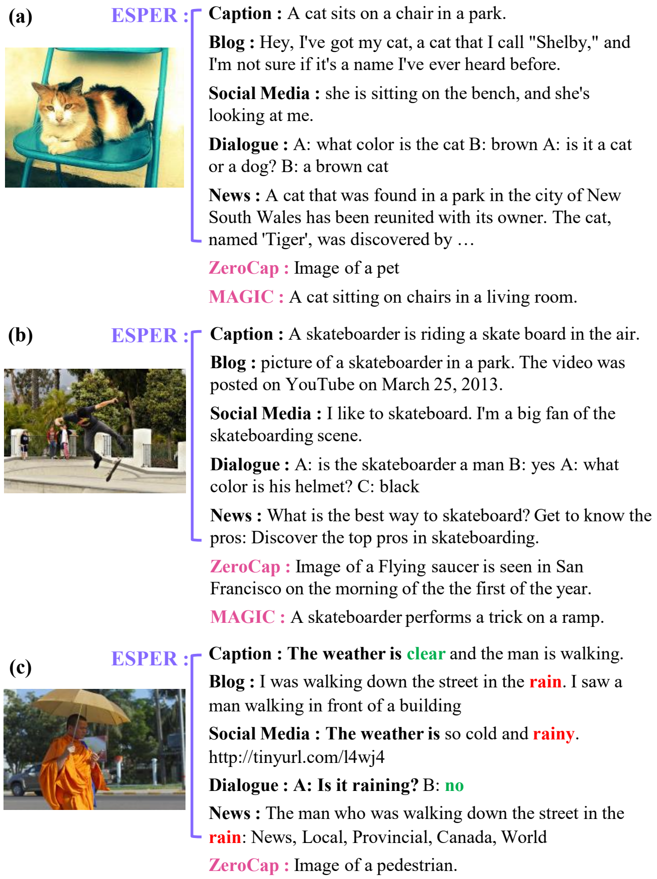
[6] Other baselines for VisualNews use the article text or keywords as inputs and hence are not directly comparable to our framework.

a COCO-supervised model, showing that ESP dataset requires domain conditioning. Also, the text-only baseline (which generates random domain-specific texts) is substantially worse, indicating the importance of visual-linguistic alignment. Finally, ESPER-Domain improves over ESPER-GPT, demonstrating the effect of explicit domain conditioning. For fine-grained results, refer to Table 6 in Appendix F.

Human evaluations on ESP dataset We conduct a human evaluation on ESPER, ZeroCap [67],[7] and ZeroCap-Domain generated descriptions as well as ground truth captions that cover six domain prompts in ESP dataset. We choose 100 random images from ESP dataset and ask English-proficient human annotators to provide a 5-point Likert-scale if the sentences: 1) are visually relevant to the image (Vis), 2) provide informative and interesting content for the prompt (Inf), 3) and sound fluent and human-like (Flu). Each sample is evaluated by three annotators using the Amazon Mechanical Turk platform. The results are shown in Figure 5. On average, ESPER provides more visually relevant and informative content than ZeroCap. We also measured Krippendorff's alpha for each category (Vis:0.61, Inf:0.50, Flu:0.31), which indicates high agreement between annotators. While ZeroCap is rated as slightly more fluent (Flu), we suspect this is due to their short text length leaving less room for grammatical errors.

3.4. Evaluation of Auditory Alignment

We extend ESPER to audio by replacing CLIP with its audio variant Wav2CLIP [74]. For the audio captioning dataset, we use the unpaired AudioCaps [27]. We follow an identical protocol as in § 3.1, but use audio modality. For baselines, we consider ZeroCap [67] and Retrieval, which first samples text using fixed prompt (*e.g.* `Sound of a`) and then retrieves ones with maximal CLIP score. ZeroCap here used Wav2CLIP as ESPER does. The results are in Table 5:[8] ESPER outperforms the baseline models. As in the visual experiments, text-only pre-training in ESPER-Domain further improves CIDEr by 5. Wav2CLIP (and early experiments with other audio encoders [18, 74, 81], also pre-trained on a classification dataset [8, 16]) provides noisier training signal for ESPER compared to image CLIP pre-trained on image caption dataset [52]. We expect this variation is not only because Wav2CLIP's datasets are relatively small [81] but also because the datasets have less rich language than image-text datasets. For samples on audio captioning, refer to Appendix I.



Caption : A cat sits on a chair in a park.

Blog : Hey, I've got my cat, a cat that I call "Shelby," and I'm not sure if it's a name I've ever heard before.

Social Media : she is sitting on the bench, and she's looking at me.

Dialogue : A: what color is the cat B: brown A: is it a cat or a dog? B: a brown cat

News : A cat that was found in a park in the city of New South Wales has been reunited with its owner. The cat, named 'Tiger', was discovered by …

ZeroCap : Image of a pet

MAGIC : A cat sitting on chairs in a living room.

Caption : A skateboarder is riding a skate board in the air.

Blog : picture of a skateboarder in a park. The video was posted on YouTube on March 25, 2013.

Social Media : I like to skateboard. I'm a big fan of the skateboarding scene.

Dialogue : A: is the skateboarder a man B: yes A: what color is his helmet? C: black

News : What is the best way to skateboard? Get to know the pros: Discover the top pros in skateboarding.

ZeroCap : Image of a Flying saucer is seen in San Francisco on the morning of the the first of the year.

MAGIC : A skateboarder performs a trick on a ramp.

Caption : The weather is **clear** and the man is walking.

Blog : I was walking down the street in the **rain**. I saw a man walking in front of a building

Social Media : The weather is so cold and **rainy**. http://tinyurl.com/l4wj4

Dialogue : A: Is it raining? B: **no**

News : The man who was walking down the street in the **rain**: News, Local, Provincial, Canada, World

ZeroCap : Image of a pedestrian.

Figure 7. ESPER Zero-shot captioning examples on various prompts. The conditioning prompt is denoted in bold(*i.e.* **"text"**). We mark visually relevant points with **green** and errors with **red**.

3.5. Qualitative Results

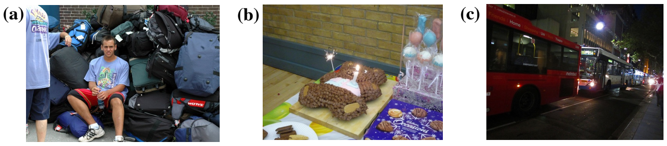
Figure 7 gives zero-shot captioning results on COCO images generated by ESPER-Domain and ZeroCap baseline [67].[9] Conditioning on both image and prefix, ESPER usually generates various visually sensible and informative captions. However, it is not perfect, *e.g.* in example (c), while the monk is holding an umbrella, it is not raining. Interestingly, ESPER is not internally consistent and, for the same image but a different prompt, correctly says the weather is clear. While ESPER sometimes generates factual dialogues, it sometimes has difficulty counting, e.g., in example *(d) Dialogue A: how many people are there B: 3.*

Figure 8 shows generation results on the "recipe" task prompt that was *not previously pre-trained as a domain prompt.* ESPER generates not only a sensible cake recipe (Fig 8 (b)), but also reasonable "recipes" even when it is not conditioned on a food image (Fig 8 (a),(c)); similar performance is observed for "My favorite poem" and "lyrics"

[7] We use a version of GPT2-base size model to generate descriptions to be comparable to our generation framework.

[8] With audio, we cannot compare to previous unpaired captioning methods [21, 34] directly, as these methods require visual object detectors.

[9] We used their public demo for qualitative results. https://replicate.com/yoadtew/zero-shot-image-to-text



(a) ESPER : (Recipe :) 1. Make a bag of luggage. 2. Put the luggage in the bag. 3. Take the suitcase to the airport. 4. Go to your hotel. 5. Get a ticket to go to a restaurant. 6. Eat at the restaurant and pay for your meal. 7. Return to where you left off. 8. Pay for the meal at your destination. 9. You're done. 10. Repeat steps 1-5 until you've paid for all your meals.

ZeroCap : (Recipe :) Camp Out Nowak Parking - The Easy Way (with Video!)

MAGIC : (Recipe :) I made a recipe for lasagna. It was very easy to make. I put the ingredients in the blender. When I was ready to serve, I took it out of the blender. My son was surprised by how much I used up the ingredients.

(b) ESPER : (Recipe :) This is a cake made with a mixture of the following ingredients: 1 cup of sugar 1/2 cup flour 1 teaspoon baking powder 1 tablespoon baking soda 1 egg 1 1⁄2 cups of milk 1 large egg yolk 1 tsp vanilla extract 1 ½ cups (1 stick) unsalted butter, softened 1½ cups sugar 2 teaspoons vanilla 1¼ cups powdered sugar.

ZeroCap : (Recipe :) cake car is made with car.

MAGIC : (Recipe :) I made a recipe for chicken noodle soup. I put it in the blender and blended it until smooth. Then I added the chicken broth and stirred it in. …

(c) ESPER : (Recipe :) 1. Place the bus in the middle of the street. 2. Turn the corner and walk to the right. 3. Walk to your left and turn right again. 4. Go to a building on your right and go to another building. 5. Turn right at the building and continue to walk. 6. At the end of the walk, you will see a door.

ZeroCap : (Recipe:) The Lazy Busway Trip Home

MAGIC : (Recipe :) I made a new recipe last night. It was for chicken noodle soup. I was trying to get it to be spicy, but my friend said it was delicious. …

Figure 8. Samples with unseen text domain prompt; **(Recipe:**).

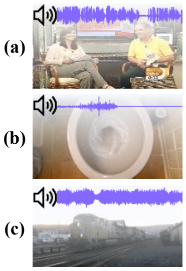
that GPT-2 can generate. In most cases, ZeroCap produces short generations and does not generalize well to custom text prompts. Yet another strong baseline, the story mode of MAGIC [64], fails to capture the visual topic and prompt.

4. Related Work

Visual-Language Pretraining. Successful vision-language models pre-trained on large-scale image-text corpora have been proposed, *e.g*. BERT-style [12] models [9, 36, 66, 80], encoder-decoder style [24, 71, 82], and contrastive models [23, 52]. Vision-text models are also extended to audio [79, 81]. TAPM [76] adapts a visual encoder and GPT with a self-supervised objective that predicts the order of the story. ESPER extend the models by training on self-generated text without any image-text pair.

Multimodal prompt tuning. Prefix tuning [37] and Prompt tuning [35] simplify finetuning large models by a fraction of the parameters. Tsimpoukelli et al. [68] adapt prefix tuning to images via maximum likelihood training a small image-to-text adapter using Conceptual Captions [61]. Like ESPER, CLIPCap [45] combines GPT + CLIP image features to generate image captions. We use the same architecture as in CLIPCap and fix GPT weights likewise, effectively following the setup of p-tuning [40].

Reinforcement learning for language tasks. In image captioning, RL helps close the gap between training and inference data [4, 55] or optimize discrete metrics directly [56]. Storytelling models employ RL to maintain coherence in the story [65] or incorporate human feedback [43]. RL is also used in goal-driven dialogue [1], interactive QA [77], and grounded generation in text games [19, 70]. Recently, Instruction GPT [48] shows RL improves the prompt-conditioning strength of pre-trained language models. To the best of our knowledge, ESPER is the first method to use multimodal rewards. While Cho et al. [10] use CLIP rewards as well, they finetune an already finetuned captioning model while ESPER builds on text-only backbones.



(a) GT : A woman and a man talking as another man talks softly and papers shuffle in the background.

ESPER : The man is talking to a woman in the background.

(b) GT : Water running from a flushed toilet.

ESPER : A toilet is flushed and flushed again.

(c) GT : Bells ring followed by humming and vibrations as a train passes while blowing a horn.

ESPER : A train is passing by a train station.

Figure 9. ESPER samples given audio inputs. Each image is the keyframe of the original video for illustration purposes. ESPER-Audio uses only audio without visual input.

5. Conclusion

ESPER combines language generation capability in a pre-trained language generator with knowledge in CLIP to align multimodal inputs to text without any supervision: we train via reinforcement learning instead of maximum likelihood training. ESPER offers strong visual alignment and fast inference speed while maintaining the text domain. We hope ESPER initiates further research on using RL for multimodal language modeling, and ESP dataset invites work on extracting diverse contexts from the same image.

6. Acknowledgements

We express special thanks to the Mosaic team members and AI2 researchers for their constructive feedback. Also, we thank Jaekyeom Kim for the comments on RL techniques. This work was supported by the Allen Institute for AI and DARPA MCS program through NIWC Pacific (N66001-19-2-4031), the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT), Artificial Intelligence Graduate School Program, Yonsei University, under Grant 2020-0-01361, and Basic Science Research Program through the National Research Foundation of Korea (NRF) (2020R1A2B5B03095585) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No.2019-0-01082, SW StarLab). We thank workers on MTurk for their contributions.

References

[1] Prithviraj Ammanabrolu, Renee Jia, and Mark O Riedl. Situated dialogue learning through procedural environment generation. In *Association for Computational Linguistics (ACL)*, 2022. 8

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 4

[3] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, 2020. 13

[4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015. 8

[5] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019. 13

[6] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 1, 2, 3, 5, 13

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 7, 18

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 8

[10] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. In *Findings of NAACL 2022*, 2022. 8

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 2, 5, 6

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 8

[13] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019. 4

[14] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*, 2020. 16

[15] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017. 14

[16] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017. 7

[17] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009. 13

[18] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022. 7

[19] Matthew Hausknecht, Prithviraj Ammanabrolu, Côté Marc-Alexandre, and Yuan Xingdi. Interactive fiction games: A colossal adventure. In *AAAI*, volume abs/1909.05398, 2020. 8

[20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. 3

[21] Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. In *EACL*, 2021. 4, 7

[22] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual Storytelling. In *NAACL-HLT*, 2016. 1

[23] 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. 8

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2, 8

[25] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 2021. 1, 2

[26] 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, 14

[27] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. 5, 7

[28] Gunhee Kim, Seungwhan Moon, and Leonid Sigal. Joint photo stream and blog post summarization and exploration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3081–3089, 2015. 1

[29] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018. 13

[30] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. *DeepMind Blog*, 2020. 3

[31] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017. 17

[32] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632, 2019. 13

[33] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. 2

[34] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards Unsupervised Image Captioning with Shared Multimodal Embeddings. In *ICCV*, 2019. 2, 4, 7, 15

[35] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 8

[36] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 8

[37] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021. 8

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755, 2014. 2, 4, 6, 13, 14, 17

[39] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *EMNLP*, 2021. 1, 2, 3, 5, 6, 13

[40] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *arXiv preprint arXiv:2103.10385*, 2021. 2, 3, 8

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 13

[42] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 5, 6

[43] Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang, Shruti Singh, Brent Harrison, Murtaza Dhuliawala, Pradyumna Tambwekar, Animesh Mehta, Richa Arora, Nathan Dass, Chris Purdy, and Mark O. Riedl. Improvisational Storytelling Agents. In *Workshop on Machine Learning for Creativity and Design (NeurIPS 2017)*, 2017. 8

[44] Alexander Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, 2016. 14

[45] Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 4, 5, 6, 8, 15

[46] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL-HLT*, 2016. 13

[47] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*, 2020. 5

[48] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 8

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 4

[50] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5, 13

[51] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, 2019. 13

[52] 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, 2, 5, 7, 8, 13, 16

[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 2, 3, 13

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 2, 3

[55] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations, ICLR*, 2016. 8

[56] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 8

[57] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035, 2019. 1, 2, 3, 5, 16

[58] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006. 13

[59] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 3

[60] Idan Schwartz. Ensemble of mrr and ndcg models for visual dialog. In *NAACL*, 2021. 1

[61] 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. 2, 8

[62] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526, 2019. 14

[63] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 3

[64] 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, 8, 15

[65] Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J Martin, Animesh Mehta, Brent Harrison, and Mark O Riedl. Controllable neural story plot generation via reinforcement learning. In *e Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. 8

[66] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *ArXiv*, abs/1908.07490, 2019. 8

[67] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*, 2021. 2, 3, 4, 5, 7, 16

[68] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Neurips*, 34:200–212, 2021. 2, 8

[69] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 4

[70] Ruoyao Wang*, Peter Jansen*, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? *arXiv preprint arXiv:2203.07540*, 2022. 8

[71] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv*, abs/2108.10904, 2021. 8

[72] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019. 3

[73] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*, 2022. 2

[74] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022. 2, 7, 18

[75] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 5, 6

[76] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. Transitional adaptation of pretrained models for visual storytelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12658–12668, June 2021. 8

[77] Xingdi Yuan, Marc-Alexandre Côté, Jie Fu, Zhouhan Lin, Chris Pal, Yoshua Bengio, and Adam Trischler. Interactive language learning by question answering. In *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2796–2813, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 8

[78] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*, 2018. 16, 17

[79] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[80] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23634–23651. Curran Associates, Inc., 2021. 8

[81] Yanpeng Zhao, Jack Hessel, Youngjae Yu, Ximing Lu, Rowan Zellers, and Yejin Choi. Connecting the dots between audio and text without parallel data through visual knowledge transfer. In *NAACL*, 2021. 7, 8

[82] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *ArXiv*, abs/1909.11059, 2020. 8