Multimodal Knowledge Alignment with Reinforcement Learning

Youngjae Yu * **Jiwan Chung** * **Heeseung Yun** **Jack Hessel**
Jae Sung Park **Ximing Lu** **Prithviraj Ammanabrolu** **Rowan Zellers**
Ronan Le Bras **Gunhee Kim** **Yejin Choi**

Allen Institute for Artificial Intelligence
Department of Computer Science and Engineering, Seoul National University
Paul G. Allen School of Computer Science, University of Washington

Abstract

Large language models readily adapt to novel settings, even without task-specific training data. Can their zero-shot capacity be extended to *multimodal* inputs? In this work, we propose ESPER (ExtraSensory PErception with Reinforcement learning) which extends language-only zero-shot models to unseen multimodal tasks, like image and audio captioning. Our key novelty is to use reinforcement learning to align multimodal inputs to language model generations without direct supervision: for example, in the image case our reward optimization relies only on cosine similarity derived from CLIP (Radford et al., 2021), and thus requires no additional explicitly paired (image, caption) data. Because the parameters of the language model are left unchanged, the model maintains its capacity for zero-shot generalization. Experiments demonstrate that ESPER outperforms baselines and prior work on a variety of zero-shot tasks; these include a new benchmark we collect+release, ESP dataset, which tasks models with generating several diversely-styled captions for each image.

1 Introduction

Zero-shot learning challenges machine learning models to make inferences for novel tasks not explicitly seen at training time. Recently, large, pretrained transformer-based models like GPT-3 (Brown et al., 2020) have achieved impressive zero-shot capabilities for a diverse set of language generation and reasoning tasks. However, models like GPT-3 only accept textual prompts as input.

In this work, we propose a new model, ExtraSensory PErception with Reinforcement learning(ESPER), that enables large language models to accept multimodal inputs like images and perform broad generation tasks over those inputs. In a zero-shot fashion, our model can generate text diverse in style and context conditioned on an image, including visual news (Liu et al., 2021a), visual dialogues (Schwartz, 2021), answers to visual questions (Antol et al., 2015; Goyal et al., 2017), visual blog-style posts (Kim et al., 2015), and visual stories (Huang et al., 2016).

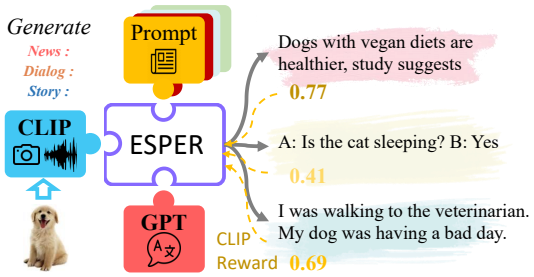

Figure 1: The intuition of ESPER, ExtraSensory PErception with Reinforcement learning. To better align knowledge in both CLIP and GPT, we give CLIP rewards on the pair of new images and self-generated text.

ESPER achieves this by combining insights from two previously disjoint lines of work: *multimodal prompt tuning*, and *reinforcement learning reward optimization*. Like prior multimodal prompt tuning work, ESPER starts from a base language-only model (e.g., GPT-2 (Radford et al., 2019)), keeps most of its parameters frozen and trains a small number of adaptor parameters to map visual features into the vocabulary space of the language model (Tsimpoukelli et al., 2021; Mokady et al., 2021; Liu et al., 2021b). *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 condi-

*denotes equal contribution

tioned on visual features. Then, parameters of a lightweight vision-to-text transformation are updated using proximal policy optimization (PPO) (Schulman et al., 2017) to maximize a similarity score computed by a secondary pretrained image-caption model, CLIP (Radford et al., 2021). The frozen language model can interpret the multimodal inputs in the same context as the initial word embedding space without additional human-annotated paired data.

A key advantage of using a reinforcement learning objective instead of a maximum likelihood objective is the maintenance of generalizability. Tsimpoukelli et al. (2021); Mokady et al. (2021) finetune their lightweight visual-to-language adapters using paired visual-linguistic datasets such as Conceptual Captions (Sharma et al., 2018) or COCO Captions (Lin et al., 2014). Because these datasets of literal descriptions cannot match the textual variety of the large-scale corpus GPT-2 is trained on, the supervised models may not generate as richly styled language or be capable of as diverse reasoning over input contexts (Kumar et al., 2022; Wortsman et al., 2022).

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 (Tsimpoukelli et al., 2021; Mokady et al., 2021); and (2) decoding-time methods (Tewel et al., 2021) 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 generation quality (e.g., 14.6 point improvement in CIDEr over Laina et al. (2019) in COCO unpaired captioning) and inference speed (e.g., $10^2\times$ speedup vs Tewel et al. (2021), which relies on per-token gradient optimization over partial decodings.)

In addition: (1) ESPER exhibits strong zero-shot adaptability on visual news (Liu et al., 2021a), visual dialogue dialogue (Das et al., 2017), and a new zero-shot multimodal generation benchmark we construct+release called ESP dataset, which tests model capacity to generate texts of different styles for the *same* image; (2) we show that ESPER can learn about audio inputs using an audio-based reward. We hope the strong performance of ESPER presented here will encourage researchers to consider RL-based training for future multimodal prompt tuning work, e.g., as a complement to max likelihood models like Flamingo-80B (Alayrac et al., 2022).

2 Method

ESPER consists of three components: 1) CLIP's non-generative image/text encoders (Radford et al., 2021);[1] 2) GPT-2 (Radford et al., 2019), a left-to-right language generator; and 3) an encoder that projects multimodal inputs into the word embedding space of GPT-2.[2] During training, CLIP and GPT-2's parameters are frozen; gradients are backpropagated through the frozen language model to train the encoder parameters. We employ reinforcement learning (specifically, PPO (Schulman et al., 2017)) to derive these gradients: the reward function is the similarity of the sampled generations to the input image, as estimated by CLIP. After RL training, we evaluate ESPER in various zero-shot scenarios.

2.1 Architecture

CLIP. Radford et al. (2021)'s Contrastive Language Image Pretrained (CLIP) encoder plays two roles in our framework: first, as a feature extractor for the input images, and second, as an alignment reward scorer between the images and the model-generated text. First, the CLIP image encoder $CLIP\text{-}I$ extracts single vector feature from the image x^i. Importantly, we do not update CLIP's parameters during training: in practice, we extract features for all images prior to training for faster execution. Second, the CLIP text encoder $CLIP\text{-}T$ is applied to text samples the model generates to support RL training; Combined with the pre-extracted image representation, this text representation is used to compute the reward function as the cosine similarity between the image and the model-generated text. While CLIP's textual representations cannot be pre-cached like the image representations because the model's generations are dynamic, because we do not backpropagate gradients to the text network this process is fast and memory-efficient to run on a GPU.

[1] While we describe image modeling here, we also experiment with audio/text encoders, specifically Wav2CLIP (Wu et al., 2022), in § 3.2 that extend ESPER to audio inputs.

[2] In principle, any models with the same APIs could be used, e.g., ALIGN (Jia et al., 2021b) could be substituted for CLIP, or T5 (Raffel et al., 2020) could be substituted for GPT-2

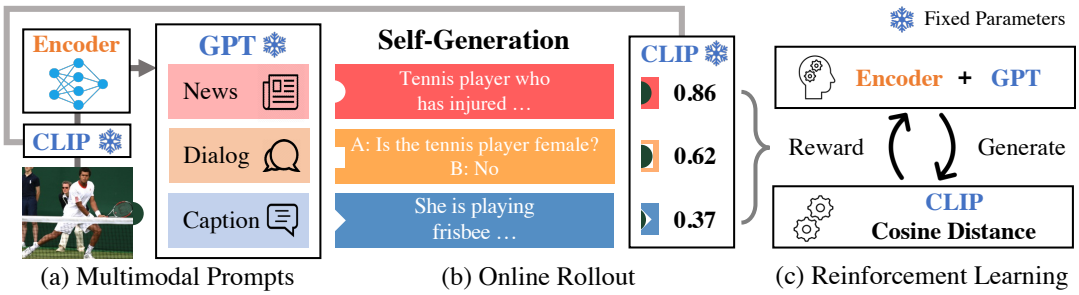

Figure 2: Illustration of the proposed model, ESPER. We use a pretrained language model (*e.g.* GPT-2 (Radford et al., 2019)) as the language generator

Encoder. The encoder F_ϕ is the only module with trainable parameters 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. The output image representations h^i work as the multimodal prompt and are concatenated to the embedded word representations. We fix the visual token length in all experiments to $k = 10$.

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

For fair comparison in later experiments, we use the same multimodal encoder architecture as CLIP-Cap (Mokady et al., 2021): a lightweight, two-layer Multi-Layer Perceptron (MLP). The first layer maps the CLIP encoding dimensions to GPT-2's dimensions and the second layer expands the single vector representation to a series of vector representations of length k. We use $tanh$ as the nonlinear activation function between these two layers. By employing a less expressive encoder architecture (than, e.g., a transformer), we aim to demonstrate that the contribution of ESPER does not rely on the structure/capacity of the encoder itself.

Pretrained Language Model. ESPER employs a pretrained deep autoregressive language model such as GPT-2 (Radford et al., 2019) as the backbone. Autoregressive language models parameterize likelihood of a text sequence y comprised of text tokens y_j with length l using autoregressive decomposition.

$$p_\theta(y) = \prod_j^l p_\theta(y_j | y_{j'<j})$$

Inspired by prompt tuning in the text-only domain (Liu et al., 2021b), we treat the encoded image vector sequence h^i as a multimodal prompt and concatenate it with the text prompt representation output by GPT-2's embedding lookup layer given previous tokens $y^i_{j'<j}$ to build the prefix for the conditioned text generation:

$$p_\theta(y^i | h^i) = \prod_j^l p_\theta(y^i_j | h^i, y^i_{j'})$$

The text prompt z can be as short as a single word token for free-form training or contain task-specific templates for further zero-shot adaption to downstream tasks.

The parameters of the language model θ are kept frozen. However, the encoder parameters ϕ are updated with the gradients calculated based on the language model parameters. Hence, we connect multimodal information to the language model without modifying the linguistic knowledge stored in the pretrained weights.

2.2 Training

Reinforcement Learning. Because CLIP does not provide per-token feedback, there is no directly differentiable way to train the encoder parameters to generate captions that CLIP would score highly, given the input image. Thus, we propose to view CLIP as a black-box model and apply reinforcement learning to minimize the embedding distance between the image context and the corresponding generated text. We use the clipped version of Proximal Policy Optimization (PPO-clip) (Schulman et al., 2017; Stiennon et al., 2020) for reward optimization. From the RL perspective, our GPT-2 generator can be viewed as a policy, which produces actions (in the form of generations) given states (in the form of text+image prompts). Our value model has the same architecture as ESPER; we use random sampling with temperature 0.7 for text generation during training.

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).

Modality Pairing Reward. The primary objective of ESPER is to align multimodal inputs to text generations. Given an input image x and the corresponding generated text y, we regard the cosine similarity between the respective CLIP features as the pairing reward.

$$r^p(x, y) \approx \frac{CLIP\text{-}I(x)}{||CLIP\text{-}I(x)||} \cdot \frac{CLIP\text{-}T(y)}{||CLIP\text{-}T(y)||}$$

The actual reward is further normalized to roughly achieve zero mean and unit variance over the course of training. In practice, we multiply the cosine similarity value with a fixed gain ($\alpha = 50$) and then add a fixed bias ($\beta = -10$).

Language Model Stability. Reward hacking can potentially occur (Krakovna et al., 2020) if the agent discovers incoherent texts that nonetheless achieve high rewards. To defend against this, we incorporate a set of 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. In addition, we found it beneficial to consider raw text-only likelihood as an additional reward. Finally, as reported in previous literature (Holtzman et al., 2020; Welleck et al., 2019), language models tend to falsely assign high probability on repetitive phrases. We introduce an explicit repetition penalty against this phenomenon. For specifics on the collection of stability rewards we apply, we refer interested readers to Appendix B.

2.3 ESPER-Style

Following previous literature on adapting language models using prompts (Gao et al., 2021), we consider a version of ESPER, where we pre-fine-tune GPT-2 with a text-only corpus alongside corresponding style prompt prefixes (*i.e.*, `"news:"`, `"story:"`). For instance, to train a news generator we present the model with a news corpus (Liu et al., 2021a) prefixed with the style prompt (`"news:"`). In practice, we finetune a single GPT on multiple styles. Note that style prompt training uses only text corpus and does not require multimodal inputs. We train these style-augmented GPT-2 generators prior to applying ESPER and provide them as backbones in place of the unconditional language models.

3 Experiments

ESP dataset. To benchmark ESPER's capability to generate diverse styles of writing from the *same* image, we collect a novel dataset: ESP dataset (Evaluation for Styled Prompt dataset). ESP dataset is a benchmark for zero-shot diverse caption generation. It comprises 4.8k captions from 1k images in the COCO Captions test set (Lin et al., 2014). We collect five different writing styles that are frequently used, namely blog, social media, instruction, story, and news, as illustrated in Figure 3. We defer the details of our data and the corresponding collection process to Section C and Section D of the Appendix, respectively.

Training. While ESPER could benefit from a more extensive and diverse set of unpaired images, for fair comparisons with the baselines, we limit our data to COCO training set images (*unpaired* with their captions). We use AdamW (Loshchilov and Hutter, 2018) optimizer ($\beta_2 = 0.999$, $\epsilon = 1e-8$) and fix the learning rate to $1e-5$ with linear decay schedule. The models are trained until there is no improvement in CLIP cosine similarity for COCO validation set images up to 50 epochs. Using a single NVIDIA A6000, and GPT-2-base/CLIP `ViT-B/32` as backbone models, ESPER needs about two days to achieve our reported evaluation scores.

ESPER Models. In addition to ESPER-Free (vanilla GPT-2 as the backbone) ESPER-Style, we experiment with ESPER-MLP, which freezes GPT-2 part of ESPER-Style and finetunes only the light MLP encoder but with supervised MSCOCO (im-

Model	Style	B@4	M	C	Time (sec/image)
Pseudo-Align (Laina et al., 2019)	✓	5.2	15.5	29.4	-
RSA (Honda et al., 2021)	✓	7.6	13.5	31.8	-
Unpaired (Laina et al., 2019)	✓	19.3	20.1	63.6	-
CLIP-Infer (Tewel et al., 2021)		2.6	11.5	14.6	65s
CLIP-Infer-Style	✓	7.0	15.4	34.5	65s
CLIP-Retrieval	✓	4.8	11.2	13.4	0.37s
ESPER-Free (GPT-2)		6.3	13.3	29.1	0.65s
ESPER-Style (GPT-2)	✓	**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 the whole time for each process needed to infer caption for an image, including image loading and feature extraction. We use greedy decoding for all results in this table.

Model	Zero-shot	B@4	M	C
CLIPCap-MLP		27.4	22.4	94.4
CLIPCap-Full		32.2	27.1	108.4
ESPER-Style	✓	21.9	21.9	78.2
ESPER-MLP		31.2	25.4	103.1
ESPER-Full		**33.1**	**27.7**	**111.1**

Table 2: Finetuning experiment in COCO Captions test split.

Model (GPT-2)	B@4	M	C
Audio Prompt + w2c	0.17	4.03	3.14
Oracle Prompt + w2c	0.80	5.34	7.07
ESPER-Audio-Free	0.36	3.05	4.68
ESPER-Audio-Style	**1.21**	**6.18**	**9.54**

Table 3: Unpaired audio captioning experiments in AudioCaps test split.

News

Model	Zero-shot	B@4	M	C
Show Attend Tell		0.7	4.1	12.2
Text-Only	✓	0.2	2.7	1.3
ESPER-Style	✓	0.8	4.4	4.6
ESPER-MLP		**1.3**	**4.8**	**15.7**

Dialog

Model	Zero-shot	NDCG	MRR	R@1
ViLBERT	✓	11.6	6.9	2.6
ViLBERT-Head		19.7	9.8	3.4
Text-Only	✓	19.3	18.3	5.7
ESPER-Style	✓	**22.3**	**25.7**	**14.6**

Table 4: Downstream task evaluation in (Visual-News (Liu et al., 2021a) test split and VisDial (Das et al., 2017) validation split. NDCG denotes Normalized Discounted Cumulative Gain, MRR Mean Reciprocal Rank and R@1 Recall at top 1. All our results on VisDial are evaluated with the official server.

age, caption) pairs and ESPER-Full trains the encoder and GPT-2 jointly with supervised MSCOCO (image, caption) pairs. All models use greedy decoding to generate descriptions at inference time.

3.1 Evaluation of Visual Alignment

We first evaluate the strength of the alignment between an input image and the generated text in ESPER. First, we consider the unsupervised task of unpaired image captioning (Feng et al., 2019). Then, we experiment with the usage of the ESPER in task transfer by comparing the trained weights with randomly initialized ones in a supervised setup. Following previous works on unpaired captioning (Feng et al., 2019; Laina et al., 2019), we split the pairing between image and caption and train them separately using ESPER for unsupervised evaluation. We split COCO Captions dataset (Lin et al., 2014) with Karpathy split (Karpathy and Fei-Fei, 2015).

3.1.1 Zero-Shot Captioning

In Table 1, we show that ESPER effectively aligns the image to text without explicitly paired data. Specifically, we compare to the state-of-the-art unpaired captioning methods (Honda et al., 2021; Laina et al., 2019) and variants of CLIP based decoding methods: CLIP-Infer (Tewel et al., 2021) that uses CLIP to guide GPT2 at inference, CLIP-Infer-Style which runs CLIP-Infer with our style-augmented GPT2 generator and CLIP-Retrieval that retrieves caption with the highest CLIP cosine similarity from the training data. According to the standard BLEU-4 (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015) automatic evaluation metrics, ESPER achieves superior performance against

previous state-of-the-art methods and CLIP based decoding algorithms. As stated in previous literature (Feng et al., 2019), we also reaffirm that style of the text affects the automatic evaluation to a great deal: ESPER-Free, which does not know COCO caption text style, falls behind ESPER-Style (which has been pretrained on unaligned COCO captions, with the prefix `caption:`).

Finally, note that the computation overhead of ESPER on inference is almost negligible compared to that of CLIP-Infer, a decoding time method (Tewel et al., 2021). On inference time, ESPER's runtime is comparable to vanilla GPT-2 alone. Only the lightweight encoder needs to run on top of GPT-2, offering fast inference speed.

3.1.2 Finetuning

As our policy network shares the same architecture with MLP-variant CLIPCap (Mokady et al., 2021), we can directly evaluate the contribution of our encoder as pretrained weights in a supervised setting. In Table 2, we show ESPER initialization bests random initialization both when updating and fixing GPT parameters. Thus, our framework can provide efficient initial alignment between two pretrained modules.

3.2 Evaluation of Auditory Alignment

We extend ESPER to another modality: audio. As an auditory counterpart of CLIP, we use Wav2CLIP (Wu et al., 2022) to score the audio-linguistic alignment during RL training, but otherwise, the setup remains the same. Here, we break the pairing in an audio captioning dataset AudioCaps (Kim et al., 2019) to evaluate unpaired audio captioning performance. We follow an identical evaluation protocol as in § 3.1, except that we only use audio as input.

In Table 3, we only report the performance of GPT-2-based baselines, as the unpaired image baselines (Laina et al., 2019; Honda et al., 2021) require object detectors and cannot be directly applied. ESPER achieves better results than baseline models, which first rollout random text samples conditioned on fixed (*e.g.* `Sound of a`) or the oracle prompts and then select ones with maximal CLIP cosine similarity. Also, the style prompt tuning positively contributes to ESPER 's performance, increasing CIDEr by 4.86. Wav2CLIP (and preliminary experiments with other audio encoders, specifically, Guzhov et al. (2022); Wu et al. (2022) which are also pretrained on an audio classification dataset (Gemmeke et al., 2017; Chen et al., 2020a)) appears to provide less accurate training signal for ESPER compared to image CLIP pretrained on large-scale image caption dataset (Radford et al., 2021). We expect this is the case not only because audio classification datasets are relatively small (Zhao et al., 2021) but also because these datasets do not offer rich natural language annotations. Still, our model can generate audio-relevant and plausible captions as described in Figure 8.

3.3 Generalization to Diverse Styles

We now experiment beyond standard image captioning setups to demonstrate ESPER's capacity to generate image-related texts of diverse styles. Here, we evaluate two styles that can be supported by existing public corpora: visual news and dialogue.

3.3.1 Visual News

VisualNews (Liu et al., 2021a) includes 1.08 million news images along with associated image captions and articles, sourced from four news sites. The captions describe the image's relevance to the news article instead of simply describing the literal image contents. For our experiments, we assign respective style prefixes per news source. For a fair comparison, we compare ESPER with models that rely only on image inputs,[3] e.g., Show Attend Tell (Xu et al., 2015), from Liu et al. (2021a). We also include the text-only style generator without visual inputs as another baseline (Text-Only).

Results are in Table 4: 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 by a wide margin in CIDEr terms of CIDEr. We attribute this difference to a combined effect of the news style and CLIP: while news consists of a myriad of proper nouns, CLIP has not been exposed to a majority of such terms. As a result, ESPER does not generate as many proper nouns as in the ground truth captions, decreasing the CIDEr score, which takes the rarity of terms into account. By finetuning the adaptor, ESPER overcomes this knowledge gap and surpasses the baselines even in the CIDEr score.

[3] Other baselines for VisualNews generate based on the article text or keywords as inputs and hence are not directly comparable to our framework.

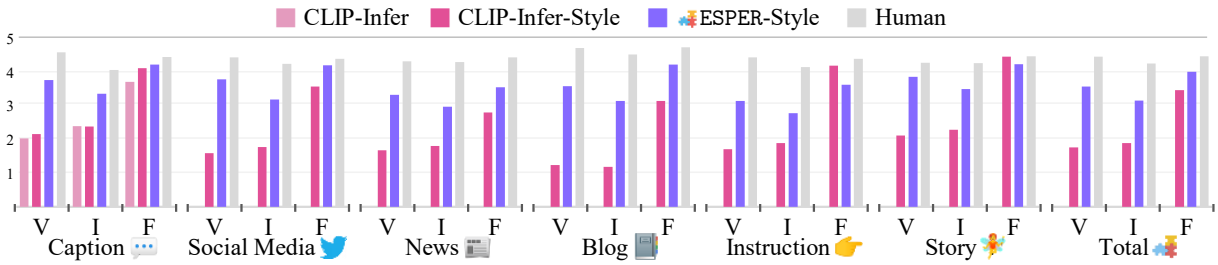

Figure 4: Human evaluation of captions for each style prompt. We take the average of 5-point Likert-scale rating from three annotators. V denotes visual relevance, I informativeness and F for fluency.

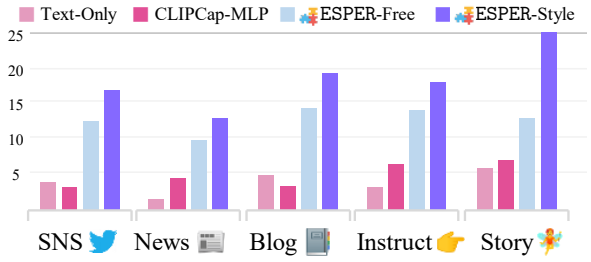

Figure 5: Evaluation on the ESP dataset. We report CIDEr in this plot.

3.3.2 Visual Dialogue

VisDial (Das et al., 2017) is a dataset of iterative dialogues conditioned on an image. Given an image and previous dialogue act, the model is asked to rank the likelihood of the 100 answer candidates. After training ESPER with the unpaired dialogue style generator backbone, we rank the answer candidates by likelihood of the answers given the image and the question. We use the validation set for evaluation for fair comparison against previously reported zero-shot baseline results (Murahari et al., 2020). The baselines consist of ViLBERT (Lu et al., 2019) and frozen ViLBERT (Lu et al., 2019) fine-tuned with a linear head.

The bottom half of Table 4 shows the VisDial dataset re-ranking results. Zero-shot ESPER improves the baselines by a margin. It even outperforms the supervised ViLBERT-Head, showing that ESPER is capable of discerning likely visual dialogues.

3.4 From One Image to Many Styles

While we observe that ESPER can generate diverse image-related texts, we still need to prove that this diversity in style is induced by text prompts. A null hypothesis is that there are identifiable and consistent features found, e.g., only in news articles. The model may have exploited this superficial relation to generating news style captions. ESP dataset from Section 3 is specifically designed to counter this hypothesis as it exhibits *multiple styled texts* for *the same image.*

Figure 5 we show that ESPER can generate diverse text depending on textual style prompts. ESPER outperforms CLIPCap-MLP (Mokady et al., 2021), a COCO-supervised baseline, demonstrating prompt-conditioned generation is necessary to handle ESP dataset. Also, the text-only baseline falls by a wide margin, indicating that the visual-linguistic alignment is as important as the text diversity. Finally, ESPER-Style improves over ESPER-Free to show the effect of explicit style conditioning. For fine-grained results, refer to Table 5 in Appendix E.

3.5 Human Evaluations on ESP dataset

We conduct a human evaluation on ESPER, CLIP-Infer[4], and CLIP-Infer-Style generated descriptions as well as ground truth captions that complete the following six prompts (`caption:`, `social media:`, `news:`, `blog:`, `instruction:`, `story:`) We choose random 100 images in ESP dataset test split 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 4. On average, ESPER provides more visually relevant and informative content in every prompt than CLIP-Infer. While CLIP-Infer has slightly more fluent (Flu) descriptions on the story and instruction prompt, we found that this is likely due to their descriptions being relatively short, thus having less room for grammatical errors.

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

(a) ESPER : **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 …
CLIP-Infer : Image of a pet

(b) ESPER : **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.
CLIP-Infer : Image of a Flying saucer is seen in San Francisco on the morning of the the first of the year.

(c) ESPER : **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
CLIP-Infer : Image of a pedestrian.

(d) ESPER : **Caption :** A group of people are sitting at a **table for food**.
Blog : I've been an audience member for a while, but I never found the right audience.
Social Media : Everyone at the conference is wearing their best suits.
Dialogue : A: dialogue: A: how many people are there B: **3** A: are they all sitting B: yes
News : Microsoft's IT team is working on a new mobile app: The company has announced that
CLIP-Infer : Image of a conference.

Figure 6: ESPER Zero-shot image captioning examples on various style prompts. The conditioning text prompt is denoted in bold(*i.e.*, **"text"**). We mark visually relevant points with **green** and errors with **red**.

3.6 Qualitative Results

Figure 6, 7 presents zero-shot captioning results on COCO images generated by ESPER-style and CLIP-Infer baseline (Tewel et al., 2021)[5].

Figure 6 shows some diverse zero-shot captions from COCO test split. Conditioning on both image and prefix, ESPER generates various visually sensible and informative captions. But Fig 6.(c) and (d) show inaccurate caption compare to the CLIP-Infer baseline. In (c), while the monk is holding an umbrella, we can deduce that it is not raining from the clear sky. However, ESPER confuses the weather condition depending on the text prompt. Also, the model suffers from false bias in visual counting. *((d) Dialogue A: how many people are there B: 3)*

Figure 7 shows generation results on the "recipe"

[5]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.
CLIP-Infer : **(Recipe :)** Camp Out Nowak Parking - The Easy Way (with Video!)

(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.
CLIP-Infer : **(Recipe :)** cake car is made with car.

(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.
CLIP-Infer : **(Recipe:)** The Lazy Busway Trip Home

Figure 7: ESPER generation results on custom task prompts, **(Recipe:)**. ESPER has never trained on recipe prompts.

(a) **(GT)** A woman and a man talking as another man talks softly and papers shuffle in the background.
(Ours) The man is talking to a woman in the background.

(b) **(GT)** Water running from a flushed toilet.
(Ours) A toilet is flushed and flushed again.

(c) **(GT)** Bells ring followed by humming and vibrations as a train passes while blowing a horn.
(Ours) A train is passing by a train station.

Figure 8: ESPER generation results on zero-shot audio captioning. Each image is the keyframe of the original video for illustration purposes. ESPER-Audio uses only audio without visual input.

task prompt that was *not previously pre-trained as a style prompt.* ESPER generate not only sensible cake recipe generation in Fig 7.(b), but also reasonable "recipe" even when it is not conditioned on a food image (Fig 7.(a),(c)); similar performance was observed for "My favorite poem" and "lyrics" that GPT-2 can generate. In most cases, CLIP-Infer generally produces short generations; and, because it wasn't designed to adapt to individual styles, it cannot as effectively generalize to custom prompts. Figure 8 additionally demonstrates that ESPER can also adapt to the audio via wav2clip rewards.

4 Related Work

Visual-Language Pretraining. Successful vision-language models pretrained on large-scale image-text corpora have been proposed, e.g., BERT-style (Devlin et al., 2019) models Tan and Bansal (2019); Chen et al. (2020b); Li et al. (2020); Zellers et al. (2021), encoder-decoder style models, Zhou et al. (2020); Wang et al. (2021); Jia et al. (2021b)

and constrastive models (Radford et al., 2021; Jia et al., 2021a). Vision-text models have additionally been extended to audio (Zhao et al., 2021; Zellers et al., 2022). TAPM (Yu et al., 2021) adapts visual encoder and GPT with self-supervised training objective that predicts causal order of the visual story.

Multimodal prompt tuning. Prefix tuning (Li and Liang, 2021) and Prompt tuning (Lester et al., 2021) simplify finetuning large models by all but a small number of parameters. Tsimpoukelli et al. (2021) adapt prefix tuning to images via maximum likelihood training a small image-to-text adapter using Conceptual Captions (Sharma et al., 2018). Like ESPER, CLIPCap (Mokady et al., 2021) combines GPT + CLIP image features to generate image caption. We use the same architecture as in CLIPCap and fix GPT weights likewise, effectively following the setup of p-tuning (Liu et al., 2021b).

Unsupervised captioning. To learn visual-linguistic relationship without paired data, previous literature draws upon pseudo-pairing retrieved with visual concept detector (Honda et al., 2021) or joint image-language embedding space (Laina et al., 2019). Most related to our work is Tewel et al. (2021) that uses CLIP image-text alignment score to guide inference of pretrained language model without further training.

Reinforcement learning for language tasks. In image captioning, RL has been used to resolve the discrepancy between training and inference data (Ranzato et al., 2016; Bengio et al., 2015) or to optimize discrete language metrics directly (Rennie et al., 2017). Storytelling models employ RL to maintain coherence in the story (Tambwekar et al., 2019) or incorporate human feedback (Martin et al., 2017). RL is also proven effective in goal-driven dialogue (Ammanabrolu et al., 2022a), interactive QA (Yuan et al., 2019), grounded generation in text games (Hausknecht et al., 2020; Wang* et al., 2022) and value alignment to human preferences (Nahian et al., 2020; Hendrycks et al., 2021; Ammanabrolu et al., 2022b). Recently, Instruction GPT (Ouyang et al., 2022) shows RL can improve prompt-conditioned generation quality of pretrained language models. To the best of our knowledge, ESPER is the first method to use multimodal reward to align images to pretrained language models; while Cho et al. (2022) used CLIP rewards as well they finetune an already trained image captioning model instead of a general large language model.

5 Conclusion

ESPER combines language generation capability in GPT-2 with multimodal knowledge in CLIP to build a diverse image-conditioned text generator: instead of maximum likelihood training, we train via reinforcement learning rewards. We note that the RL objective we consider can be used in conjunction with multimodal prompt tuning (Tsimpoukelli et al., 2021) and zero-shot captioning with CLIP guidance (Tewel et al., 2021).

Future work includes:

1. enhancing ESPER so that it can simultaneously maximize rewards for multiple modalities (Image, Audio, OCR, Motion in video, etc.);
2. scaling up the CLIP and GPT-2 backbones to larger variants; and
3. exploring the utility of ESPER as a data augmentation tool for multimodal reasoning tasks.

6 Acknowledgements

We express special thanks to the Mosaic team members and AI2 researchers who gave feedback on this project. Also, we express our gratitude for the helpful comments by Jaekyeom Kim on reinforcement learning techniques. This work was supported by the Allen Institute for AI and DARPA MCS program through NIWC Pacific (N66001-19-2-4031). SNU members are supported by 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 all our workers on MTurk for their contributions to our project.

References

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

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

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajizhirzi, and Yejin Choi. 2022b. Aligning to social norms and values in interactive narratives. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

Satanjeev Banerjee and Alon Lavie. 2005. 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.

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

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

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. 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.

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

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

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

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

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

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

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

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

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics (ACL)*.

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

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*.

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

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

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021. What would jiminy cricket do? towards agents that behave morally. *NeurIPS*.

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

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

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. 2016. Visual Storytelling. In *NAACL-HLT*.

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

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

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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.

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

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

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

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. 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.

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

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards Unsupervised Image Captioning with Shared Multimodal Embeddings. In *ICCV*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. 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.

Xiang Lisa Li and Percy Liang. 2021. 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.

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. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

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

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021a. Visual news: Benchmark and challenges in news image captioning. In *EMNLP*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. GPT understands, too. *arXiv preprint arXiv:2103.10385*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

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. 2017. Improvisational Storytelling Agents. In *Workshop on Machine Learning for Creativity and Design (NeurIPS 2017)*.

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

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*.

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

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

Md Sultan Al Nahian, Spencer Frazier, Mark Riedl, and Brent Harrison. 2020. Learning norms from stories: A prior for value aligned agents. In *AAAI/ACM Conference on AI, Ethics, and Society*, page 124–130, New York, NY, USA. Association for Computing Machinery.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. 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.

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

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

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

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

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

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

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

Idan Schwartz. 2021. Ensemble of mrr and ndcg models for visual dialog. In *NAACL*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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.

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

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

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

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

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2021. Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*.

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

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

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

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

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

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. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*.

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

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

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. 2021. 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.

Xingdi Yuan, Marc-Alexandre Côté, Jie Fu, Zhouhan Lin, Chris Pal, Yoshua Bengio, and Adam Trischler. 2019. 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. Association for Computational Linguistics.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. 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)*.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *Advances in Neural Information Processing Systems*, volume 34, pages 23634–23651. Curran Associates, Inc.

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

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

A Language Model Backbones

Figure 9: Annotation interface of ESP dataset.

ESPER**-Free.** We use GPT-2-base (Radford et al., 2019) as the language model backbone for all experiments. Since GPT-2 does not have a special start-of-sentence token, we provide a random single token as an initial text prompt to start generation on. This initial token is sampled from GPT-2 vocabulary with sampling weight computed with token frequency.

ESPER**-Style.** As summarized in Section 2.3, we finetune GPT-2 on text-only corpus with style prompts to prepare the style generator backbones. The style prompts and corresponding text corpus sources include:

- `caption`: COCO Caption (Lin et al., 2014)
- `social media`:
  - Sentiment140 (Go et al., 2009)
  - MDID (Kruk et al., 2019)
  - TweetEval (Barbieri et al., 2020)
- `news`: GoodNews (Biten et al., 2019)
- `blog`: Blog Authorship (Schler et al., 2006)
- `instruction`: WikiHow (Koupaee and Wang, 2018)
- `story`:
  - ROCStories (Mostafazadeh et al., 2016)
  - TimeTravel (Qin et al., 2019)

For visual news generation, we use different style prompt per news source (`bbc:`, `guardian:`, `usa today:`, `washington post:`) to reflect writing style differences between media in VisualNews dataset (Liu et al., 2021a).

B Language Model RL Training

KL Divergence. By constraining KL divergence between the online policy and the initial language model, we aim to maintain salience of the generated text. Here, we simply optimize the difference between the log likelihood of the online policy and the initial policy for each token generated.

Reference Entropy. To constrain deviation from text generation capability, we first compute text-only log likelihood using either the pretrained text style generator or the vanilla language model. Then, we penalize the model whenever the text-only negative log likelihood of a generated token exceeds a predefined threshold $\tau_e = \frac{70}{l}$, where l is the length of the generated sequence. We take inverse of the difference between negative log likelihood and threshold and optimize it as a reward. In practice, we further scale this reward with fixed gain $\alpha_e = 0.1$.

		Social Media			News			Blog			Instruction			Story			Total		
Model	Prompt	B	M	C	B	M	C	B	M	C	B	M	C	B	M	C	B	M	C
Text-Only	✓	0.2	3.7	3.9	0.0	2.2	1.6	0.3	4.1	4.9	0.0	4.0	3.3	0.3	4.7	5.9	0.1	3.7	3.9
ClipCap-MLP		0.0	3.9	6.8	0.0	4.8	7.5	0.3	4.0	6.6	0.3	4.2	7.6	0.0	4.3	7.3	0.1	4.2	7.2
	✓	0.2	3.0	3.3	0.2	3.9	4.5	0.0	2.9	3.4	0.5	4.8	6.5	0.0	4.4	7.1	0.2	3.8	5.0
ESPER-Free	✓	**0.6**	5.6	12.5	0.6	5.5	9.9	**0.7**	6.2	14.4	**0.7**	5.6	14.1	0.6	5.7	13.0	0.6	5.7	12.8
ESPER-Style	✓	**0.6**	**5.8**	**16.9**	**0.7**	**5.7**	**13.0**	**0.7**	**6.7**	**19.2**	**0.7**	**5.7**	**18.0**	**1.2**	**7.5**	**25.0**	**0.8**	**6.3**	**18.4**

Table 5: Style-wise experiment results on ESP dataset. B denotes Bleu-4 score.

	Caption			Social Media			News			Blog			Instruction			Story			Total		
Model	Vis.	Inf.	Flu.	Vis.	Inf.	Flu.	Vis.	Inf.	Flu.	Vis.	Inf.	Flu.	Vis.	Inf.	Flu.	Vis.	Inf.	Flu.	Vis.	Inf.	Flu.
Clip-Infer	1.98	2.34	3.62	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Clip-Infer-Style	2.11	2.33	4.01	1.56	1.73	3.48	1.64	1.76	2.73	1.21	1.16	3.06	1.67	1.85	**4.09**	2.07	2.23	**4.35**	1.72	1.85	3.38
ESPER-Style	**3.67**	**3.27**	**4.12**	**3.69**	**3.11**	**4.10**	**3.24**	**2.90**	**3.46**	**3.49**	**3.06**	**4.12**	**3.06**	**2.71**	3.53	**3.76**	**3.41**	4.13	**3.48**	**3.08**	**3.91**
Human	4.47	3.96	4.34	4.32	4.14	4.28	4.21	4.19	4.33	4.60	4.41	4.62	4.32	4.04	4.28	4.17	4.16	4.36	4.35	4.15	4.36

Table 6: Human evaluation of captions for each style prompt. We take the average of 5-point Likert-scale rating from three annotators. Vis. denotes visual relevance, Inf. informativeness and Flu. for fluency.

Repetition Penalty. This reward penalizes the model for generating repeated n-grams. Given GPT tokenizer, we count repeated (1, 2, 3)-grams. Specifically, we subtract the number unique of n-grams from that of all n-grams to count repetitions. Then we compute a weighted sum of the n-gram repetition counts and scale the combined score with fixed gain $\alpha_r = 0.025$ and bias $\beta_r = 0$.

C Details on ESP dataset

There are multiple ways to describe an image depending on the context and intent of the author. We refer to these multiple methods as *styles*. Previous works focus on the sentiment of a caption like positive & negative (Mathews et al., 2016), romantic & humorous (Gan et al., 2017), and various personalities (Shuster et al., 2019). However, style does not solely depend on sentiments and emotions: it comprises every choice of text type, structure and vocabulary used to convey intended meaning of the writer. As intention of a writer depends on where the one's interest lies, different information of the same visual cue would be illustrated on each style of writing.

For example, consider an image of a boy with a bow tie singing as part of a choir on a stage. While this image may have been uploaded by the singer's sibling with a caption like "go bro, love the bowtie!", a local news article about the same concert might instead write: "the choir's performance on August 17th went off without a hitch." Because different styles of writing may focus on different aspects of an image, and styles may not be fully translatable via text-only operators such as text style transfer, e.g., "go bro, love the bowtie!" doesn't mention anything about a choir performance.

We thus collect ESP dataset to explore broad range of text styles conditioned on the same image. Using Amazon Mechanical Turk, we ask the annotators to write captions relevant to an image while following writing styles mentioned above. An image cost about $0.3 to annotate, which translates to $7-28 of payment per a work hour depending on the proficiency of the worker. The average length of ESP dataset is 28.4 words (2.3 sentences), and the collected captions are filtered with respect to their adherence to given images and styles.

D ESP dataset Collection Process

We use Amazon Mechanical Turk to collect captions as shown in Figure 9. For images in COCO Captions test set with respect to Karpathy split, we randomly select images with one to five annotated objects to select images with salient but not noisy context. We ask the annotators to write sentences that are relevant to the image while following the mentioned writing style. We provide examples from well-constructed datasets as references, as listed in text corpus sources of Appendix A. We ask the annotators to write no less than 30 words, but for writing styles with shorter text like social media and news, we lower the bar from 30 to 10 words. We also regularly monitor the collection so that only the workers with fluency and understanding of style can participate in the process. In total, 189 workers participated in the collection process. The collected dataset is filtered by manually verifying whether the captions are relevant to given images and styles.

E ESP dataset Experiment Details

We compare ESPER against three baselines in this experiment. The first is a text-only baseline. We use the pretrained text style generator with random sampling to generate the candidate texts. The rest two baselines (Mokady et al., 2021) are trained on a caption supervision dataset (COCO captions) and share the same architecture as our ESPER. As the supervised baselines are not intended for prompt conditioning, we report evaluation results both with and without the style prompts for them. When not using the style prompts, we fix the prompt to "Image of a", following the recommended approach in literature (Mokady et al., 2021). For fair comparison against the baselines trained with supervised dataset of limited length (ClipCap-MLP), we truncate all text including the ground truth captions to the first 20 byte-pair tokens with GPT tokenizer. Note that all compared methods share the same tokenization scheme as the vanilla GPT2 and hence the truncation does not favour any specific approach.

We report the evaluation results in Table 5. For clarification, the scores in Table 5 include and expand upon the summarized results in Figure 5. ESPER shows flexible adaptability to each style without being exposed to any paired image-text data of the given styles. On the other hand, the supervised baselines exhibit limited generalizability to diverse text styles even when conditioned on style prompts. The total score is computed as the mean over metrics of each style, without considering sample size difference.

F Additional Qualitative Samples

In Figure 10- 11, we displays ESPER generation results conditioned on custom prompts such as (*This is my favorit poem*) or (*I created lyrics*). The conditioning text prompt is denoted as bold font enclosed with parenthesis (*i.e.*, **"(text)"**). To qualitatively emphasize the randomness of our results, we provided the model with progressively growing prompts.

(a)  (b)  (c) 

(a) (This is my favorite poem :)

I love the way the pitcher swings his bat. I like the fact that the ball is so quick to come out of his hand. I like how he can hit the baseball with his left hand, but he's not a great hitter with the right.

(This is my) first time seeing a baseball game. I've never seen a game like this before. The game was so fast and exciting. It was like watching a movie.

(This) is a photo of a baseball player in a uniform. urlLink This is the player's name. urlLink

(I) was in the batter's box, and I was trying to hit the ball. I hit it, but it went out of the park.

(b) (This is my favorite poem :)

I am a giraffe, I have a heart, and I love the giraffes. I am the heart of the world, , the sun of my life, my heart is the sky, and the earth is my home, and the moon is mine

(This is my) favorite giraffe in the world. I love it. It's so cute. I love the way it looks. The way the ears are. And the tail. That's the best part.

(This) is a picture of a giraffe in a zoo. The giraffes are in the wild. They are not in captivity. The girafen are the largest of the girurans. It is estimated that there are about 1,000 of them in the wild. The giraffe is the only mammal that can walk on two legs.

(I) was in the zoo with my family, and I was very excited to see the giraffe. I had never seen a giraffe before, but I knew it was a big animal. It was so big, it looked like a lion.

(My favorite lyric :)" I love the giraffes, they are the best in the world. They are so cute, and they have such a sweet personality. "

(c) (This is my favorite poem :)

I have a craving for pizza. I've been craving pizza for a long time. It's been a while since I had a pizza, but I'm finally getting one. And I love it.

(This is my) first time eating pizza. I'm not sure if I'll ever eat pizza again.

(This) is a very good pizza. I love the crust and the cheese. The sauce is good, but I would have preferred a little more pepper. It's not bad, just not my favorite.

(I) am not sure if I'm going to eat a pizza today, but I think I'll eat it anyway. I've been eating pizza for years, and I love it.

Figure 10: ESPER generation results conditioned on custom prompts.

(a)  (b)  (c) 

(a) (I created lyrics:)

I'm a baby elephant, I have a big head, and I am a little elephant.

I'm not a child, but I do have big ears. I can hear you

(I created) this blog to share the news of elephants and other animals in the area. I hope you enjoy it. I am a member of the Elephant Conservation Society of America (ECSA)

(I) was very surprised to find that the zoo had not yet opened a new exhibit. We had been told that it would be open in the fall,

(b) (I created lyrics:)

I created lyrics: I don't know, no, it's so beautiful in the ocean, so serene, and I'm so happy to be here.

(I created) a number of Facebook posts about the boat that I was visiting, and I began writing them. The first post was about how I had been told that the ship was going to be sunk

(I), along with other senior citizens, drove the small boats to the port of the morning. There, they were met by a group of men in civilian clothes, who told them that they had been ordered

(c) (Question :)

So, what is the purpose of surfing? I think it's to get out of the house and get away from the world. I think that's the main reason why I surf

(Lyrics:) I'm surfing in the water. I can't see the ocean.

I'm riding in a boat. It's dark. The waves are high. My heart is pounding.

(Recipe:) urlLink This is a very simple surfboard trick. You can use it to make a surf board that is very easy to build. It is made of a piece of wood,

(Why) is it so hard to surf in the water?

The answer is that the waves are so big that they are impossible to see. The waves can be seen from the shore

Figure 11: ESPER generation results conditioned on custom prompts.