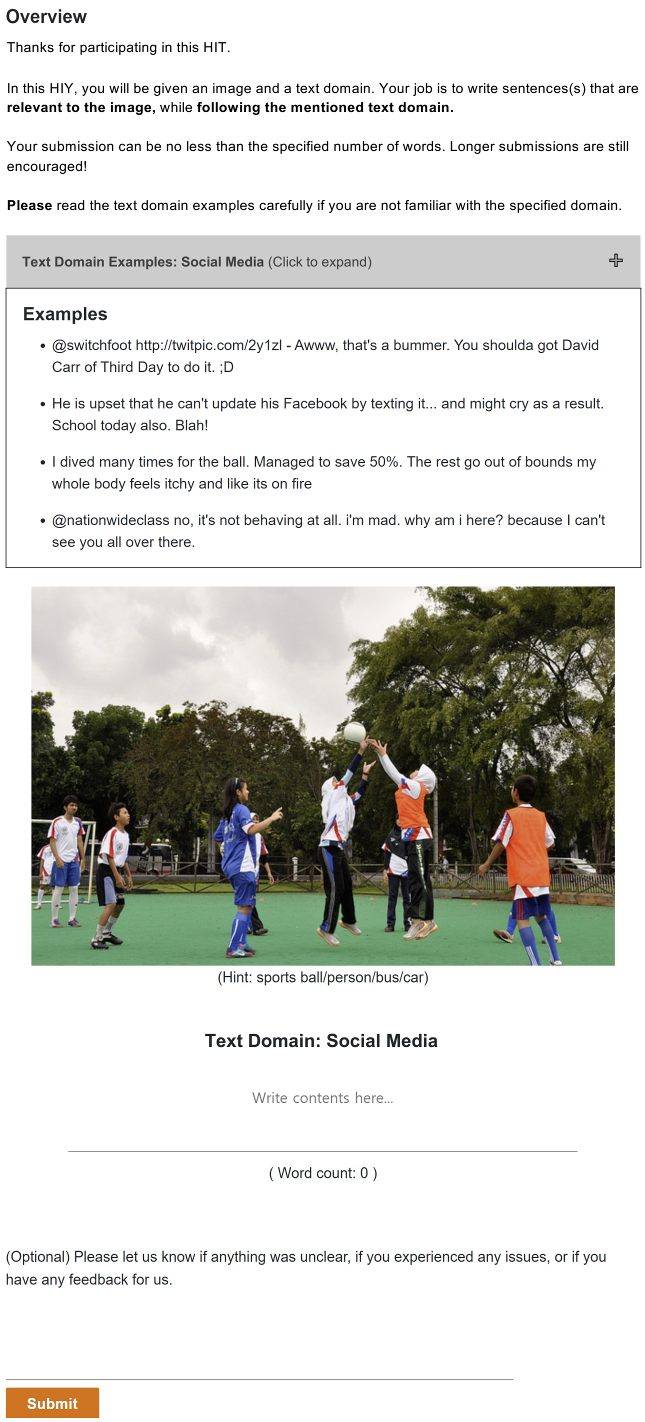
A. Language Model Backbones



Overview

Thanks for participating in this HIT.

In this HIY, you will be given an image and a text domain. Your job is to write sentences(s) that are **relevant to the image,** while **following the mentioned text domain.**

Your submission can be no less than the specified number of words. Longer submissions are still encouraged!

Please read the text domain examples carefully if you are not familiar with the specified domain.

Text Domain Examples: Social Media (Click to expand)

Examples

- @switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
- He is upset that he can't update his Facebook by texting it... and might cry as a result. School today also. Blah!
- I dived many times for the ball. Managed to save 50%. The rest go out of bounds my whole body feels itchy and like its on fire
- @nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.

(Hint: sports ball/person/bus/car)

Text Domain: Social Media

Write contents here...

(Word count: 0)

(Optional) Please let us know if anything was unclear, if you experienced any issues, or if you have any feedback for us.

Figure 10. Annotation interface of ESP dataset.

ESPER-GPT. We use GPT-2-base [53] 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-Domain. As summarized in § 2.3, we prepare the domain-specific text generators by finetuning GPT-2 on text-only corpus with domain prompts. These domain-specific models are trained with conventional teacher forcing and aim at prompting the open-ended GPT to focus on a specific "style" (e.g. news or blogs). The domain prompts and corresponding text corpus sources include:

- caption: COCO Caption [38]
- social media:
  - Sentiment140 [17]
  - MDID [32]
  - TweetEval [3]
- news: GoodNews [5]
- blog: Blog Authorship [58]
- instruction: WikiHow [29]
- story:
  - ROCStories [46]
  - TimeTravel [51]
- commonsense:
  - COMET [6]
  - VCG [50]

When training commonsense data, we concatenate the context sentence, relation and output sentence with special indicators in-between to build a single text string input (*e.g.* [SOS] fire hydrant ObjectUse [GEN] get water for fire [EOS]).

For visual news generation, we use different prompt per news source (bbc:, guardian:, usa today:, washington post:) to reflect writing style differences between media in VisualNews dataset [39].

B. Training Details

The only trained part of ESPER is the multimodal encoder with 8M parameters when using the frozen GPT-2-base backbone (124M parameters). The multimodal encoder projects the CLIP [52] image embedding to 10 continuous representations, which are then inserted as token embeddings to the language model. In training ESPER, we use AdamW [41] 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. For automatic evaluation of the generated text, we use three metrics: BLEU-4 based on 4-gram precision, METEOR using unigram precision and recall and CIDEr aiming to capture human consensus. Implementation of the metrics follows COCO evaluation code (https://github.com/tylin/coco-caption).

C. Language Model RL Training

Reinforcement Learning Our value model shares the same architecture as ESPER policy model; we use random sampling for text generation during training with the same architecture as the policy model P to train them in actor-critic fashion. The value model uses the generated text and image as inputs, and it produces scalar values for each token that indicate the expected return of the current state. At each iteration $k+1$, the text samples y are generated given the multimodal prompt x using the policy model from the previous step P_{ϕ_k}. We use random sampling with a low temperature (0.7) for text generation during training.

$$y \sim P_{\phi_k}(\cdot|x)$$

Then, the advantage value A can be calculated for the value model V_k from the previous step k as follows:

$$A^k(x,y) = V^k(x,y) - r(x,y)$$

Given the context x and the corresponding sample y, we search for ϕ to maximize the PPO-clip objective:

$$\min\left(\frac{P_\phi(y|x)}{P_{\phi_k}(y|x)}A^k(x,y), g(\epsilon, A^k(x,y))\right)$$

$$g(\epsilon, A) = \begin{cases}(1+\epsilon)A, A \geq 0\\(1-\epsilon)A, A < 0\end{cases}$$

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 domain-specific language 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 $\alpha_e = 0.1$.

Repetition Penalty. We penalize 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 $\beta_r = 0$.

D. 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 type of descriptions as *domains*. Previous works focus on the sentiment of a caption like positive & negative [44], romantic & humorous [15], and various personalities [62]. However, text domain 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 one's interest, different information of the same visual cue would be illustrated on each domain 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 writing domain may focus on different aspects of an image, writings of different domain may not be fully inter-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 domains conditioned on the same image. Using Amazon Mechanical Turk, we ask the annotators to write captions relevant to an image while following text domains 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 text domain.

E. ESP dataset Collection Process

We use Amazon Mechanical Turk to collect captions as shown in Figure 10. For images in COCO Captions test set [38] with respect to Karpathy split [26], we randomly select images with one to five annotated objects to select

		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-GPT	✓	**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-Domain	✓	**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 6. Domain-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.
ZeroCap	1.98	2.34	3.62	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ZeroCap-Domain	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-Domain	**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 7. Human evaluation of captions for each domain prompt. We take the average of 5-point Likert-scale rating from three annotators. Vis. denotes visual relevance, Inf. informativeness and Flu. for fluency.

images with salient but not noisy context. We ask the annotators to write sentences that are relevant to the image while following the mentioned text domain. 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 text domains 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 text domain 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 text domains.

F. ESP dataset Experiment Details

We compare ESPER against three baselines in this experiment. The first is a text-only baseline. We use the pre-trained domain-specific language generator with random sampling to generate the candidate texts. The rest two baselines [45] 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 domain prompts for them. When not using the domain prompts, we fix the prompt to ”Image of a”, following the recommended approach in literature [45]. For fair comparison against the baselines trained with a 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 GPT-2 and hence the truncation does not favor any specific approach.

We report the evaluation results in Table 6. For clarification, the scores in Table 6 include and expand upon the summarized results in Figure 6 of the main paper. ESPER shows flexible adaptability to each domain without being exposed to any paired image-text data of the given domain. On the other hand, the supervised baselines exhibit limited generalizability to diverse text domains even when conditioned on domain prompts. The total score is computed as the mean over metrics of each domain, without considering the sample size difference.

Model	Encoder	B@4	M	C
Unpaired [34]	-	19.3	20.1	63.6
MAGIC [64]	-	12.9	17.4	49.3
ESPER	Linear	21.9	21.9	78.4
ESPER	Transformer	19.4	20.2	68.2

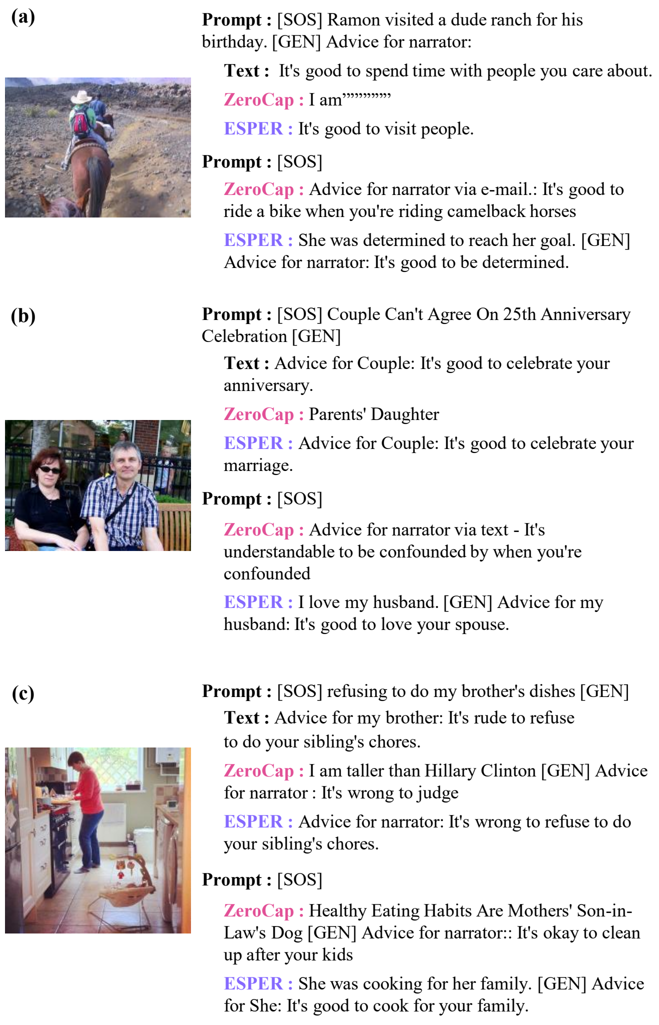
Table 8. Alternative visual encoder architecture experiment in unpaired COCO-Captions test split. We use the linear encoder in other experiments.

G. Alternative Encoder

Here, we demonstrate that the improvement of ESPER is not confined to a specific architecture. Table 8 shows the performance ESPER when combined with an alternative transformer encoder introduced in CLIPCap [45] as our base linear encoder does as well. Even with an alternative encoder, our ESPER shows stable generation quality and outperforms the baselines. However, the transformer variant falls behind a simpler architecture of linear encoder. We attribute this phenomenon to the larger exploration space in the parameters: in the training process, we observed that the transformer encoder requires more training steps and samples to reach the same performance as the linear encoder.

Model	SOS			Head		
	B@4	M	C	B@4	M	C
Retrieval [52]	5.9	14.6	17.1	-	-	-
Text-Only	1.1	9.5	0.4	3.1	8.0	41.8
ZeroCap [67]	0.0	0.0	0.0	0.0	0.8	3.6
ESPER	8.2	14.8	19.8	3.0	7.6	40.2

Table 9. Commonsense inference experiments in the validation split of SWAG dataset [78]. SOS and Head each denote giving only [SOS] token or the full head sentence as the text input. Retrieval has no Head-prompted result as most of visually-aligned context is in the head, rendering retrieval given head meaningless.



(a)

Prompt : [SOS] Ramon visited a dude ranch for his birthday. [GEN] Advice for narrator:

Text : It's good to spend time with people you care about.

ZeroCap : I am’’’’’’’’’’

ESPER : It's good to visit people.

Prompt : [SOS]

ZeroCap : Advice for narrator via e-mail.: It's good to ride a bike when you're riding camelback horses

ESPER : She was determined to reach her goal. [GEN] Advice for narrator: It's good to be determined.

(b)

Prompt : [SOS] Couple Can't Agree On 25th Anniversary Celebration [GEN]

Text : Advice for Couple: It's good to celebrate your anniversary.

ZeroCap : Parents' Daughter

ESPER : Advice for Couple: It's good to celebrate your marriage.

Prompt : [SOS]

ZeroCap : Advice for narrator via text - It's understandable to be confounded by when you're confounded

ESPER : I love my husband. [GEN] Advice for my husband: It's good to love your spouse.

(c)

Prompt : [SOS] refusing to do my brother's dishes [GEN]

Text : Advice for my brother: It's rude to refuse to do your sibling's chores.

ZeroCap : I am taller than Hillary Clinton [GEN] Advice for narrator : It's wrong to judge

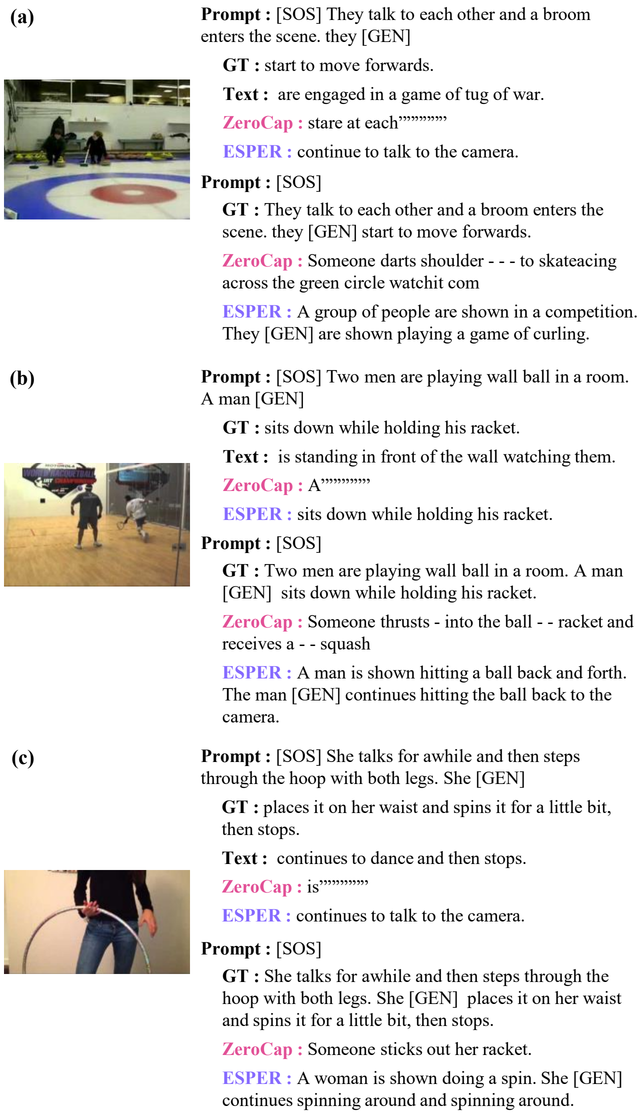
ESPER : Advice for narrator: It's wrong to refuse to do your sibling's chores.

Prompt : [SOS]

ZeroCap : Healthy Eating Habits Are Mothers' Son-in-Law's Dog [GEN] Advice for narrator:: It's okay to clean up after your kids

ESPER : She was cooking for her family. [GEN] Advice for She: It's good to cook for your family.

Figure 11. ESPER generation results in Social Chemistry 101 dataset [14].



(a)

Prompt : [SOS] They talk to each other and a broom enters the scene. they [GEN]

GT : start to move forwards.

Text : are engaged in a game of tug of war.

ZeroCap : stare at each’’’’’’’’’’

ESPER : continue to talk to the camera.

Prompt : [SOS]

GT : They talk to each other and a broom enters the scene. they [GEN] start to move forwards.

ZeroCap : Someone darts shoulder - - - to skateacing across the green circle watchit com

ESPER : A group of people are shown in a competition. They [GEN] are shown playing a game of curling.

(b)

Prompt : [SOS] Two men are playing wall ball in a room. A man [GEN]

GT : sits down while holding his racket.

Text : is standing in front of the wall watching them.

ZeroCap : A’’’’’’’’’’

ESPER : sits down while holding his racket.

Prompt : [SOS]

GT : Two men are playing wall ball in a room. A man [GEN] sits down while holding his racket.

ZeroCap : Someone thrusts - into the ball - - racket and receives a - - squash

ESPER : A man is shown hitting a ball back and forth. The man [GEN] continues hitting the ball back to the camera.

(c)

Prompt : [SOS] She talks for awhile and then steps through the hoop with both legs. She [GEN]

GT : places it on her waist and spins it for a little bit, then stops.

Text : continues to dance and then stops.

ZeroCap : is’’’’’’’’’’

ESPER : continues to talk to the camera.

Prompt : [SOS]

GT : She talks for awhile and then steps through the hoop with both legs. She [GEN] places it on her waist and spins it for a little bit, then stops.

ZeroCap : Someone sticks out her racket.

ESPER : A woman is shown doing a spin. She [GEN] continues spinning around and spinning around.

Figure 12. ESPER generation results in SWAG dataset [78].

H. Fusing Text-Only Commonsense Data

Dataset. We evaluate ESPER against three text-only commonsense datasets. We flatten the inputs and outputs of each dataset into a single string using the template ([SOS] Head Relation [GEN] Tail [EOS]). Basic information and the flattened text samples of each dataset are as follows:

- **ATOMIC** [57]: commonsense knowledge
  - ATOMIC connects the cause and effects of everyday events using nine types of If-Then relations.
  - *e.g.* [SOS] *PersonX calls the police xIntent* [GEN] *PersonX wants to report a crime* [EOS]
- **Social Chemistry 101** [14]: social and moral norm
  - Given an everyday situation, Social Chemistry

101 annotates free-text rules-of-thumbs for acceptable social and moral behaviors.

  - *e.g.* [SOS] *Asking my boyfriend to stop being friends with his ex* Advise for *Narrator*: [GEN] *It's not right to tell another person who to spend time with* [EOS]

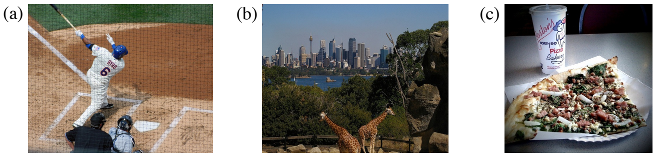
- **SWAG** [78]: grounded commonsense inference
  - SWAG unifies natural language inference with commonsense reasoning by suggesting a physically possible follow-up event to the given previous event.
  - *e.g.* [SOS] *The woman is now blow drying the dog. The dog* [GEN] *is placed in the kennel next to a woman's feet.* [EOS]

Visual Alignment. There is no ground-truth image-caption pair label in our selection of text-only datasets. However, a subset of SWAG is derived from ActivityNet Captions dataset [31], which has paired video labels. We take the middle frame from each video and use it as input to the visually-aligned methods (ESPER, ZeroCap) for evaluation.

For ATOMIC, we build an artificial evaluation dataset with image-caption pairs from COCO captions dataset [38] as follows: i) to ensure visual alignment, we take image captions from COCO captions dataset. ii) then, we extract commonsense outputs with the text-only COMET model pretrained on ATOMIC using COCO-captions as text inputs, and iii) concatenate the captions and COMET outputs to get the evaluation set and filter out with human supervision. After filtering, we leave 2000 visually aligned samples.

Experiments. Table 9 shows that ESPER generates plausible commonsense natural language inference pairs given the corresponding images. Also, the ESPER-trained model performs on par with the base text model when the oracle query (head) is given, showing that ESPER conserves commonsense knowledge in the text backbone. Interestingly, ZeroCap builds degenerate text when used with a SWAG-pretrained text backbone. We suspect that as the SWAG dataset is relatively small in size, the text backbone is overfitted and sends imprecise gradient signals to the ZeroCap decoder. This shortcoming of the decoding-time algorithm further exemplifies the robustness of ESPER in fusing weaker text backbones. For experiments on ATOMIC, refer to § 3.2.1 in the main paper.

Qualitative Samples. Figure 11 shows samples generated using ESPER on Social Chemistry 101 dataset. As the dataset does not have visual grounding annotation, we take images from COCO dataset. We use two types of text prompts; SOS, which uses only the [SOS] token as the input, and Head, which takes the given head sentence as an additional input. To ensure acceptable visual alignment in the heads, we use CLIP to retrieve image-head pairs with cosine similarity higher than 0.28.



(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.

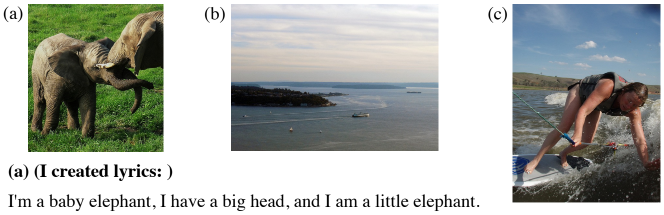
(b) (This is my favorite poem :)
I am a giraffe, I have a heart, and I love the giraffes. I am the heart of the world, , the sun of my life, my heart is the sky, and the earth is my home, and the moon is mine
(This is my) favorite giraffe in the world. I love it. It's so cute. I love the way it looks. 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 13. ESPER generation results conditioned on custom prompts.

Given the head sentence, the ESPER-trained model generates similar text to the text-only model, preserving commonsense knowledge in the text backbone. On the other hand, ZeroCap breaks the predefined text syntax in sample (b) and even fails to construct a coherent sentence in sample (c). Furthermore, when given the SOS-only prompt ESPER keeps the text syntax relatively intact than ZeroCap.

In Figure 12, we show ESPER samples on SWAG dataset. Here, we exploit the ground-truth video-text alignment to achieve image-head pairs, as explained previously. ESPER shows meaningful visual alignment while keeping the text style of the pretrained backbone. However, ESPER builds rather dull text than the ground-truth and tends to repeat the content of the head in the tail sentence (b and c). We attribute this weakness to an artifact of using CLIP-derived reward without strong regularization: a simple shortcut to increasing such reward is repeating words or phrases already known to be associated with the image context. We believe these examples illustrate the need for an RL objec-



(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 14. ESPER generation results conditioned on custom prompts.

tive better tailored for a multimodal environment.

I. Additional Qualitative Samples

Figure 9 of the main paper shows samples from zero-shot generation with ESPER given audio inputs and audio caption text generator. ESPER-Audio captures the objects (*e.g*. woman and man, toilet and train) and the main action (*e.g*. talking, flushing and passing by) of the auditory scene. However, the generated captions are shorter and less informative compared to the ground truth captions. We attribute this difference to the nature of Wav2CLIP [74]: the audio encoder is pretrained on a audio classification dataset [8] and lacks understanding of fine-grained audio-linguistic alignment.

In Figure 13- 14, we display ESPER generation results conditioned on custom prompts such as (*This is my favorite 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.