Learning to Imagine: Visually-Augmented Natural Language Generation

Tianyi Tang[1,4], **Yushuo Chen**[1], **Yifan Du**[1],
Junyi Li[1,3], **Wayne Xin Zhao**[1,4 ✉], and **Ji-Rong Wen**[1,2,4]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]DIRO, Université de Montréal
[4]Beijing Key Laboratory of Big Data Management and Analysis Methods
steventianyitang@outlook.com chenyushuo1999@foxmail.com
lijunyi@ruc.edu.cn {yifandu1999,batmanfly}@gmail.com

Abstract

People often imagine relevant scenes to aid in the writing process. In this work, we aim to utilize visual information for composition in the same manner as humans. We propose a method, **LIVE**, that makes pre-trained language models (PLMs) **L**earn to **I**magine for **V**isually-augmented natural language g**E**neration. First, we imagine the scene based on the text: we use a diffusion model to synthesize high-quality images conditioned on the input texts. Second, we use CLIP to determine whether the text can evoke the imagination in a posterior way. Finally, our imagination is dynamic, and we conduct synthesis for each sentence rather than generate only one image for an entire paragraph. Technically, we propose a novel *plug-and-play fusion layer* to obtain visually-augmented representations for each text. Our vision-text fusion layer is compatible with Transformer-based architecture. We have conducted extensive experiments on four generation tasks using BART and T5, and the automatic results and human evaluation demonstrate the effectiveness of our proposed method. We will release the code, model, and data at the link: https://github.com/RUCAIBox/LIVE.

1 Introduction

Natural language generation (NLG) is a fundamental technique for supporting a variety of downstream applications (Li et al., 2022b; Zhao et al., 2023), *e.g.,* text summarization, story generation, and data-to-text generation. As the mainstream NLG approach, pre-trained language models (PLMs) can produce human-like text under the guidance of input conditions. Despite their success, these models are pre-trained on the text-only corpora, and they cannot well capture visually-grounded semantics, *e.g.,* visual commonsense (Ilharco et al., 2021), making it difficult to achieve desired results when visual knowledge is required.

✉ Corresponding author

To improve the generation capacity of PLMs, existing work has widely explored various methods to incorporate visual knowledge into models, which can be roughly divided into two lines of research. The first line designs specific visually-enhanced training tasks such as continual pre-training on text-image data (Cho et al., 2021) or knowledge distillation with vision-language models (Dai et al., 2022). However, these methods usually perform well only on multimodal generation tasks (*e.g.,* visual question answering) but not text generation tasks, due to the semantic disparity across modalities (Tan and Bansal, 2020). As the second line, several studies retrieve or synthesize images related to the input and then fuse the image representations into PLMs (Wang et al., 2022b; Zhu et al., 2022). However, they simply treat the input as a whole (even for long texts) for retrieving or synthesizing related images, which cannot sufficiently leverage fine-grained visual semantics.

Considering the above issues, we are motivated by the process of human writing where they have the ability to imagine relevant scenes from the contexts in their minds. These visual scenes convey related experiences in the world that can inspire the human's writing (Bisk et al., 2020; Popham et al., 2021). By imitating such behavior, we consider NLG as a writing process of a human, where the input text is conditioned on a set of dynamically "*imagined scenes*", *i.e.,* synthesized images.

To this end, in this paper, we propose a novel approach, **LIVE**, that enables PLMs to **L**earn to **I**magine for **V**isually-augmented natural language g**E**neration. Different from previous methods, our augmentation approach is relevant, selective, and dynamic. To be *relevant*, we utilize the state-of-the-art text-to-image model, Stable Diffusion (Rombach et al., 2022), to synthesize realistic images for fine-grained semantic units (*i.e.,* sentences). Compared to the retrieval-based approach, our method can generate more relevant, diverse images that

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9468–9481

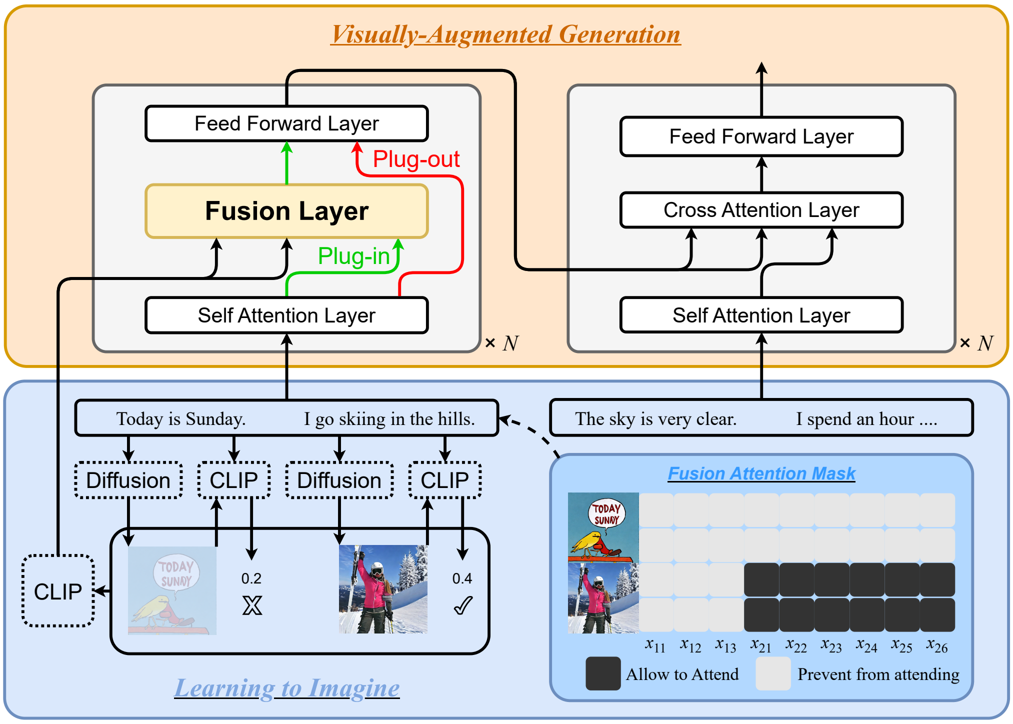

Figure 1: The overall illustration of our proposed approach **LIVE**, consisting of the text-related image generation and the plug-and-play vision-text fusion layer. The fusion attention mask means that the first sentence x_1 lacks visuality and will skip the fusion layer (green flow), while the second sentence x_2 has high visuality, and each word x_{2i} of x_2 will attend to the synthesized image to obtain visually-augmented text representations (red flow).

may not exist in real-world image databases. To be *selective*, we evaluate the degree to which the text's meaning can be visualized in an image and only invoke the use of synthesized images when it is actually needed. To be *dynamic*, we synthesize images for each sentence in the *input* text so that the visual knowledge is more fine-grained compared to a single image for the whole input. In order to deeply fuse the visual knowledge of synthesized images, we propose a *plug-and-play vision-text fusion layer* for Transformer-based models. We also design specific mechanisms to support efficient text-image cross-attention and enable the controllability of the use of visual knowledge.

Our contributions are summarized as follows:

• We propose a new approach, called LIVE, to learning to use synthesized images to improve natural language generation, imitating the process of human writing.

• We propose a plug-and-play vision-text fusion layer to incorporate visual knowledge and obtain visually-augmented text representations.

• We verify the effectiveness of our approach with BART and T5 on four text generation tasks: LIVE consistently outperforms these PLMs, with an average improvement ratio of 2.44%.

2 Related Work

Pre-Trained Models. In recent years, large-scale pre-training has achieved remarkable success and has become the dominant technique in the NLP community (Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020; Zhao et al., 2023). Pre-trained on massive text corpora, models can learn contextualized representations that include both linguistic and world knowledge (Jiang et al., 2021). Since PLMs are trained with pure text corpora without connection to the visual world, vision-language pre-training (VLP) leverages image-text pairs to learn cross-modal representations (Gan et al., 2022; Su et al., 2020; Li et al., 2020; Radford et al., 2021). It has been discovered that VLP models have more visual knowledge than PLMs (Ilharco et al., 2021), however, they cannot perform well on text-only tasks such as language understanding (Yun et al., 2021). In this work, we mainly focus on incorporating visual knowledge to enhance the performance of natural language generation tasks based on existing text-only models.

Visually-Augmented Language Learning. Considering the leakage of visual knowledge in language models, many researchers attempt to enhance text-only tasks with visual information, which is known as visually-augmented (aided or grounded) language learning. Vokenization (Tan and Bansal, 2020) and iACE (Lu et al., 2022) propose to treat contextualized-related images as vokens and pre-train a text model to predict them for fusing visual information. Similarly, VidLanKD (Tang et al., 2021) extends finite image vokens to diverse video frames and employs a knowledge distillation method to acquire visual knowledge. The subsequent works leverage CLIP (Radford et al., 2021) as the vision source to integrate visual information into PLMs via CLIP output embeddings (Wang et al., 2022b; Guo et al., 2022) or knowledge transfer methods (Dai et al., 2022; Jin et al., 2022). The majority of these works can outperform PLMs on language understanding tasks. As for natural language generation tasks, researchers mainly focus on finding suitable images and fusing the visual representations into text-only models using a shallow module. Some works apply generation models, such as GAN-based models (Long et al., 2021; Zhu et al., 2022) and VAE-based models (Fang and Feng, 2022), to synthesize (latent) images, while Liang et al. (2021), Shen et al. (2021), and Su et al. (2022) propose to employ contextualized text embeddings to retrieve relevant images. In our work, we utilize the superior diffusion model (Rombach et al., 2022) to synthesize high-quality images and propose a plug-and-play vision-text fusion layer to deeply integrate visual knowledge into PLMs and obtain visually-augmented text representations.

Multimodal Language Generation. Multimodal language generation aims to produce fluent and coherent text based on the input text or image. Different from unimodal language generation, the additional image serves as the background for generation. Multimodal language generation includes tasks such as image caption (Lin et al., 2014), visual question answering (Zhang et al., 2016), multimodal machine translation (Elliott et al., 2016), multimodal text summarization (Jangra et al., 2021), visual dialog (Das et al., 2017), and visual story telling (Huang et al., 2016). However, the construction of these datasets requires costly manual annotation, which hinders their widespread application. In contrast, we do not require text-image pairs as input and instead utilize Stable Diffusion (Rombach et al., 2022), a text-to-image model, to synthesize images for input texts.

3 Method

3.1 Task Formulation

Natural language generation (*a.k.a.,* text generation) aims to capture the semantic mapping relation from an input text $\mathcal{X} = \langle x_1, ..., x_k, ..., x_m \rangle$ to an output text $\mathcal{Y} = \langle y_1, ..., y_k, ..., y_n \rangle$, where x_k and y_k denote the k-th sentences of the input and output texts, respectively. In this paper, we focus on the task of *visually augmented natural language generation (VA-NLG)*. Following prior works (Zhang et al., 2020; Wang et al., 2022b), VA-NLG further assumes text-related image data can be obtained to help text generation. Here, we consider a generalized way (*e.g.,* retrieval and synthesis) to obtain the related images with an image augmenter $\mathcal{F}$, where $\mathcal{F}$ takes as input a sentence x (or a text) and outputs an image i_x related to x: $\mathcal{F}(x) \rightarrow i_x$.

The goal of VA-NLG is to generate readable and plausible output texts $\mathcal{Y}$ based on input texts $\mathcal{X}$ and image augmenter $\mathcal{F}$, which is formally defined as:

$$\mathrm{P}(\mathcal{Y}|\mathcal{X}) = \prod_{k=1}^{n} \mathrm{P}(y_k|\mathcal{X}, y_{<k}; \mathcal{F}), \tag{1}$$

where $y_{<k}$ denotes previously-generated sentences.

With this formulation, there are two key issues for this task: (1) how to design the image augmenter to obtain potentially useful images, and (2) how to use the augmented images for improving text generation. Considering the two issues, we propose **LIVE**, a general approach to augmenting NLG tasks with related images, with sentence-level image synthesis via text-to-image diffusion model (Section 3.2) and plug-and-play vision-text fusion for using augmented images (Section 3.3).

3.2 Text-Related Image Generation

Although it is intuitive to augment PLMs with visual images, it is challenging to obtain appropriate and helpful images for given texts. Some previous work (Zhang et al., 2020; Tan and Bansal, 2020) utilizes retrieval-based methods to search images from text-image databases, such as MS COCO (Lin et al., 2014). However, these static image resources are limited in both *quantity* and *content*, which is likely to result in inaccurate image retrieval.

Synthesizing Relevant Images. To circumvent the limitation of static image resources, we instead propose to automatically generate images for given texts by leveraging text-to-image generation models. In contrast to previous works that utilize GAN-based (Esser et al., 2021) or auto-regressive (Wang et al., 2022a) generation models, we use the state-of-the-art Stable Diffusion model (Rombach et al., 2022), a probabilistic diffusion model guided by CLIP-encoded input text representations, to synthesize high-quality images. With Stable Diffusion, we can flexibly perform image generation based on different text units. Here, we consider *sentences* as synthesis units, which contain a moderate amount of information for an image. Compared with the previous work that synthesize a single image for the whole input, our sentence-level generation is more fine-grained. It is inspired by the writing behavior of people: one would switch the imagined scenes for different sentences.

For each input sentence x_k, we apply Stable Diffusion to synthesize its corresponding creative image i_{x_k}. Equipped with the acceleration method of DDIM (Song et al., 2021), Stable Diffusion is able to synthesize photographic images normally in 50 steps (Rombach et al., 2022). In practice, we empirically find that using a 25-step synthesis can usually lead to a decent performance in our task (see Section 5.4 for more analysis about the diffusion quality and efficiency).

Evaluating the Text Visuality. Although the generation-based method is flexible to produce images on various topics, not all texts can inspire the generative model to generate meaningful images, such as the rule text *"ACL 2023 requires all papers to have a clear discussion of limitations"*. Only texts with visually rich content can be associated with images. Previous work usually synthesizes or retrieves images without considering the visuality of texts, tending to incorporate irrelevant or noisy images. However, it is difficult to directly measure the visuality of a text. As a compromise, we estimate the similarity score in a posterior way between a sentence x_k and a synthesized image i_{x_k} using CLIP (Radford et al., 2021):

$$\gamma = \text{CLIP}(x_k, i_{x_k}) \in [-1, 1] \,. \tag{2}$$

CLIP is a vision-language model pre-trained on a massive amount of text-image pairs using contrastive learning which excels at evaluating the similarity between text and image. In our work, we manually set a threshold value θ. If γ exceeds the threshold value, the text is considered to have high visuality; otherwise, we consider that the text has weak visuality and discard the synthesized image. We will discuss the influence of θ in Section 5.3.

3.3 Plug-and-Play Vision-Text Fusion

After synthesizing relevant images for given texts, we study how to leverage visual images for improving text generation. Instead of using VLP models, we aim to fuse the visual knowledge into a PLM-based backbone, since text generation is essentially a language modeling task. To enhance the cross-modality fusion, we propose a plug-and-play vision-text fusion module to obtain deeply-fused visually-augmented text representations.

Vision-Text Fusion for PLMs. Our fusion module is a plug-and-play attention layer for Transformer-based (Vaswani et al., 2017) models, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). We insert the fusion layer after the self-attention layer in the encoder. Our fusion layer is a layer-wise cross-attention module to augment the word representations with visual information. In particular, for a sentence x_k and the corresponding synthesized image i_{x_k}, we first utilize CLIP to encode the image into patch representations $\mathbf{I}_k \in \mathbb{R}^{p \times d}$. Then, we feed the sentence into the Transformer model and obtain the output representation $\mathbf{S}_{k,l}$ for the self-attention sub-layer in the l-th layer of the encoder. Finally, we pass $\mathbf{S}_{k,l}$ to our l-th plug-and-play fusion layer to obtain the visually-augmented text representations:

$$\mathbf{F}_{k,l} = \begin{cases} \text{FusionLayer}_l(\mathbf{S}_{k,l}, \mathbf{I}_k, \mathbf{I}_k), & \gamma \geq \theta \\ \mathbf{S}_{k,l}, & \gamma < \theta \end{cases}, \tag{3}$$

where γ is the similarity score computed in Equation 2, and FusionLayer_l conducts multi-head attention on the query, key, and value matrices, followed by residual connection and layer normalization. Here, we introduce γ to control whether a generated image will be used or not.

In general, such a fusion layer can be applied to various Transformer-based PLMs and LLMs. Note that each sentence attends to no more than one image, as depicted in the attention matrix in Figure 1. Compared to simply concatenating images and text as input (Liang et al., 2021), our cross-attention-based mechanism is more efficient while maintaining performance (see Section 5.2). Besides, our fusion is more controllable and can achieve fine-grained cross-attention. For example, we can choose only nouns to be attended with images since they contain more visual information (see Section 5.2).

3.4 Optimization

In order to achieve decent performance, we can pre-train the key component of our approach, *i.e.*, the fusion layer (Section 3.3), with text-image paired datasets. Specially, we collect the image caption datasets MS COCO (Lin et al., 2014), Flickr30k (Plummer et al., 2015), CC3m (Sharma et al., 2018), and Visual Genome (Krishna et al., 2017) as text-image pairs, and utilize the caption text to synthesize images using Stable Diffusion to enrich the pre-training pairs. In this way, we can obtain 9 million text-image pairs in total. Then, we apply image-based denoising autoencoding as the pre-training objective, which teaches the model to recover the caption based on a noisy text. Such a pre-training strategy can make the fusion layer better map the visual knowledge into text space.

Next, we describe the overall optimization process of our approach. During pre-training, we freeze the PLM backbone and only pre-train the fusion layer; therefore, if we plug-out the fusion layer, the PLM retains its original language generation ability. The fusion layer is a lightweight module and has 18M parameters for $\text{BART}_{\text{BASE}}$ (140M). During fine-tuning, we utilize Stable Diffusion and CLIP models to synthesize images and compute similarity scores. These operations can be done offline for efficiency, and the diffusion and CLIP models will not be updated. We only need to fine-tune the whole PLM as usual, in addition to the small pre-trained fusion layer.

4 Experiment

4.1 Experimental Setup

4.1.1 Dataset

We conduct experiments on four text generation datasets with diverse tasks and domains:

• E2E (Novikova et al., 2017) is a data-to-generation task with the aim of converting multiple input meaning representations into fluent texts.

• CommonGen (Lin et al., 2020) requires the model to generate a coherent and reasonable text given a collection of common concepts.

• SAMSum (Gliwa et al., 2019) is a dialogue summarization dataset that evaluates the model's summary and dialogue understanding abilities.

• ROCStories (Mostafazadeh et al., 2016) consists of five-sentence stories, and we utilize the first sentence as input to generate the remaining four.

The details of the statistics and license of each dataset are listed in Table 1. For each dataset, we utilize NLTK[1] to tokenize the input texts into sentences, except that we treat each key-value pair in the input as a sentence for the E2E dataset.

Dataset	#Train	#Valid	#Test	License
CommonGen	67,389	993	–	MIT
E2E	42.061	547	630	CC BY-SA 4.0
ROCStories	176,688	9,816	4,909	N/A
SAMSum	14,732	818	819	CC BY-NC-ND 4.0

Table 1: The statistics and licenses of datasets.

4.1.2 Evaluation Metrics

We adopt five automatic metrics, namely BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), and Distinct (Li et al., 2016), to compare the performance of different methods. BLEU, ROUGE, and CIDEr compute the n-gram overlap between the candidate text and the reference text(s). SPICE further takes semantic meaning into consideration. Distinct mainly evaluates the diversity of the generated texts and is always used in open-ended generation tasks, such as story generation. We also conduct the human evaluation in Section 5.5.

4.1.3 Baseline Models

We utilize two commonly used text generation PLMs, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), as text-only baselines. We further compare them to two multimodal VLP models:

• BLIP (Li et al., 2022a) uses a multimodal mixture of encoder-decoder with the objectives of text-image contrast, text-image matching, and language modeling on bootstrapped text-image pairs.

• OFA (Wang et al., 2022a) unifies text and image modalities using a unified architecture and multi-task sequence-to-sequence learning. In addition, we consider a variant and attempt to use OFA with only text, denoted by OFA *w/o* image.

We integrate our LIVE framework with BART and T5, and consider the following visually-augmented methods as comparisons:

• VL (Cho et al., 2021) adds visual embeddings for the original BART and T5 and conducts continued pre-training on text-image pairs.

• iNLG (Zhu et al., 2022) guides the PLM with the machine-generated image as the visual prefix.

[1] https://www.nltk.org/

Methods	E2E			CommonGen			SAMSum			ROCStories	
	B-4	R-L	ME	B-4	CIDEr	SPICE	R-1	R-2	R-L	B-1	D-4
BLIP	45.05	54.35	34.84	13.30	5.84	18.62	22.54	4.07	20.56	28.29	66.93
OFA	67.20	69.18	45.12	29.34	15.48	30.79	47.42	23.20	43.45	31.70	68.16
OFA *w/o* image	67.63	69.08	45.19	29.54	15.46	30.84	48.12	23.33	43.81	32.51	70.99
BART	67.38	69.57	45.04	30.30	16.05	31.16	49.92	25.55	45.61	32.98	76.77
VL-BART	68.53	69.57	45.17	29.51	15.19	29.54	45.02	20.22	40.83	32.76	76.32
iNLG-BART	64.71	67.19	43.14	29.80	15.80	30.78	50.75	26.20	46.36	33.25	50.87
LIVE-BART	**69.24**	**70.59**	**45.60**	**31.47**	**16.55**	**31.89**	**51.31**	**26.67**	**47.08**	**33.46**	**79.98**
T5	66.54	68.02	44.71	26.70	15.66	30.96	49.27	**25.30**	45.18	33.14	75.11
VL-T5	66.96	70.09	44.66	27.29	15.31	29.78	49.91	24.95	45.20	33.07	75.09
LIVE-T5	**68.34**	**71.11**	**46.09**	**27.94**	**15.84**	**31.36**	**49.99**	25.16	**45.84**	**33.22**	**77.28**

Table 2: The results of four text generation tasks. B, R, ME, and D are short for BLEU, ROUGE, METEOR, and Distinct, respectively. The best results are highlighted in **bold**. These setups and abbreviations are the same below.

Since iNLG does not offer a T5 version, we can only combine it with BART for comparison.

4.1.4 Implementation Details

For all baselines, we utilize the base versions of PLMs, *i.e.,* $\text{BART}_{\text{BASE}}$, T5_{BASE}, $\text{BLIP}_{\text{BASE}}$, and OFA_{BASE}, which have a comparable number of parameters to ensure a fair comparison. For BLIP, OFA, VL-BART, and VL-T5, we provide the same synthesized image as our method, and we fine-tune them similarly to how they perform VQA tasks. For iNLG, we utilize its official implementation[2].

As for our method, we employ Stable Diffusion v1.4 with half precision[3] to synthesize images in 25 timesteps for efficiency. Then, we adopt CLIP-ViT-B/32 to judge the similarity between text-image pairs and extract image features. We empirically set the threshold value $\theta = 0.27$. After extraction, an MLP layer is appended to project the image representation into the text space and obtain an image representation $\boldsymbol{I}_i \in \mathbb{R}^{50\times768}$. The aforementioned operations can be performed offline for efficiency.

In the pre-training stage of our fusion layer, we mask 50% of the input text with span lengths drawn from a Poisson distribution with $\lambda = 3.5$ for BART and force the model to recover the input with the image. As for T5, we split the caption into two parts and train the model to generate the second part using the first part and the image. We pre-train the fusion layer with a batch size of 384, optimize BART using AdamW (Loshchilov and Hutter, 2019) with a constant learning rate of 1×10^{-5}, and optimize T5 using Adafactor (Shazeer and Stern, 2018) with a learning rate of 1×10^{-3}.

[2]https://github.com/VegB/iNLG

[3]https://huggingface.co/CompVis/stable-diffusion-v1-4

In the fine-tuning stage, we tune the entire model, including the PLM backbone and the fusion layer. We set the batch size to 32 and employ the same optimizer and learning rate as in pre-training. We optimize the model using cross-entropy sequence-to-sequence loss with a label smoothing factor (Szegedy et al., 2016) of 0.1. During inference, we choose the checkpoint with the highest validation metric score for generation. During generation, we apply beam search with a beam size of 5 for E2E, CommonGen, and SAMSum, while utilizing the nucleus sampling with $p = 0.9$ and $t = 0.7$ for ROCStories.

All the experiments are conducted using the text generation library TextBox (Tang et al., 2022) on NVIDIA GeForce RTX 3090 24GB GPUs using Ubuntu 20.04.1 SMP. *All these hyper-parameters are identical for our method and baselines.*

4.2 Experimental Results

Based on the results in Table 2, we can find that:

Firstly, the results of multimodal models (*i.e.,* BLIP and OFA) cannot achieve satisfactory results when compared with text-only models (*i.e.,* BART and T5) on pure text tasks. This finding further proves the existence of semantic disparity (Tan and Bansal, 2020) across modalities of generation tasks. OFA without images even outperforms OFA with images slightly, which indicates that images may be a burden for text generation tasks when the fusion method is not appropriate.

Secondly, the visually-augmented methods (*i.e.,* VL-BART, VL-T5, and iNLG) can achieve superior performance than their base PLMs on certain tasks but cannot achieve overall improvement on all tasks. A major reason might be that they synthesize only one image for each input without considering

Methods	0.1%			0.3%			1%			3%		
	B-4	R-L	ME	B-4	R-L	ME	B-4	R-L	ME	B-4	R-L	ME
BART	50.58	57.95	32.37	56.18	62.34	36.02	62.11	66.38	39.34	65.25	68.15	42.18
iNLG-BART	28.40	53.89	25.98	39.15	58.63	30.05	48.66	62.12	33.31	61.74	65.75	38.05
LIVE-BART	**51.67**	**60.41**	**33.06**	**60.87**	**64.32**	**38.22**	**63.31**	**67.00**	**40.30**	**65.99**	**69.08**	**43.00**

Table 3: The few-shot experiments on the E2E dataset.

its relevance and sentence-level semantics.

Finally, our LIVE method can outperform all baselines on all four text generation tasks. Equipping BART with our LIVE method, LIVE-BART can outperform its text-only counterpart BART by 2.80% in ratio. LIVE can also work with T5, yielding an average improvement of 2.08%. These automatic results demonstrate the effectiveness and compatibility of our text-related image generation approach and plug-and-play fusion layer.

5 Further Analysis

In this section, we conduct various experiments to test the efficacy of our methods. The tuning details are identical to those introduced in Section 4.1.4.

5.1 Few-Shot Results

We investigate the performance of our LIVE methods in a low-resource situation. We keep 0.1%, 0.3%, 1%, and 3% of the training set for the E2E dataset. For each split, we choose five independent groups to decrease the randomness. From the results in Table 3, we can observe that our methods remarkably boost the performance under few-shot settings compared with baselines, especially in extreme situations (0.1% and 0.3%). We assume that synthesized images can provide visual knowledge as a supplement when training data is scarce.

5.2 Ablation Study

To examine the effectiveness of the different factors of our LIVE methods, we conduct four groups of experiments for ablation. The results are reported in Tables 4 and 5. First, we can see that the *pre-training* of the vision-text fusion layer is beneficial.

Second, we replace the *image augmenter* $\mathcal{F}$ Stable Diffusion with two variants: a text-image retriever CLIP (Radford et al., 2021) and a text-to-image synthesizer VQGAN (Esser et al., 2021). We can find that the synthesis-based methods are superior to the retrieval-based ones since they can generate relevant images which may not exist in a static database. Compared with VQGAN, Stable Diffusion can synthesize high-quality images and provide more visual knowledge for text generation.

	B-4	R-L	ME
LIVE-BART	69.24	70.59	45.60
w/o pre-training	68.02	69.72	45.33
Image augmenter			
CLIP	65.70	68.65	44.63
VQGAN	67.13	69.42	45.15
Fusion method			
Concatenation	67.30	69.37	45.12
Self-attention	68.08	69.72	45.28

Table 4: Ablation analysis on the E2E dataset. The experiments with different image augmenters and fusion methods are conducted without pre-training.

Image source	B-4	R-L	ME
Sent-level (Ours)	69.24	70.59	45.60
Doc-level	68.25	70.24	45.26
Selective sent-level	69.30	70.62	45.69
Word-level	67.67	69.58	45.36

Table 5: Further analysis on the different granularities of different image synthesis strategies.

Third, we investigate the *fusion method* of visual representations and make two variants of our cross-attention-based fusion. "Concatenation" means to concatenate the image representations and the encoder output as the input for the decoder, while "Self-attention" means to concatenate the image representations and the text representations as the input for the encoder. The results indicate that the deep fusion of text and vision representations is beneficial and the cross-attention-based method and self-attention-based method are comparable, which is consistent with Gan et al. (2022). Thus, we utilize cross-attention as the fusion method because it is more efficient and controllable.

Finally, we explore our dynamic and controllable fusion layer. To be dynamic, we synthesize one image for each sentence in the input (denoted as "Sent-level") and attempt two variants that synthesize one image for the whole document ("Doc-level") or each word in the document ("Word-level"). The re-

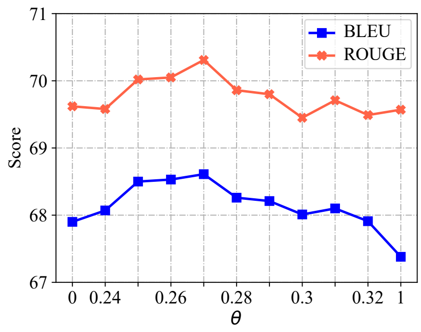

Figure 2: Varying the similarity threshold value θ.

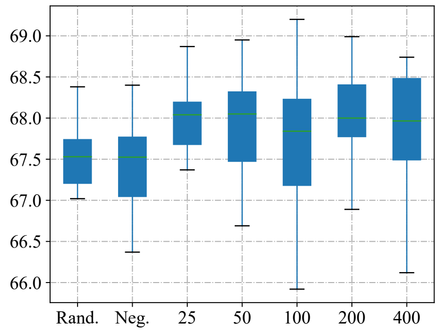

Figure 3: Varying the number of diffusion steps.

sults prove the effectiveness of our sentence-level synthesis compared with previous method (Zhu et al., 2022) that only generates one image for the input. However, too many images actually lead to poor performance. In addition, we investigate a fine-grained cross-attention based on sentence-level synthesis ("Selective sent-level"). We only make noun words visually-augmented and make the other words skip the fusion layer. The results show that the fine-grained fusion may be promising, and we leave it for future work.

5.3 Model Sensitivity *w.r.t.* the Similarity Threshold Value θ

In Section 3.2, we set a threshold value θ to measure the text visuality. Here, we investigate the model's performance when θ varies. If $\theta = 0$, all the sentences will be visually-augmented. If $\theta = 1$, all the sentences will not be visually-augmented, and it degenerates to text-only BART. As shown in Figure 2, LIVE-BART with $\theta = 0.27$ achieves the best performance, and we find that 0.27 is close to the median of text visuality scores, *i.e.,* nearly half of the sentences will be augmented and the others will not be. Therefore, we set $\theta = 0.27$ for our LIVE methods in experiments.

Datasets	LIVE+BART wins	Ties	BART wins
E2E	29%	56%	15%
CommonGen	24%	58%	18%
SAMSum	40%	34%	26%
ROCStories	48%	11%	41%

Table 6: Human evaluation on four generation tasks.

5.4 Model Sensitivity *w.r.t.* the Synthesized Images

In this subsection, we first demonstrate that visual information is truly favorable for text generation. Following the previous works (Zhang et al., 2020), we replace the image representations with random noise or utilize the input text as a negative prompt to synthesize irrelevant images. The results in Figure 3 further prove the necessity of visual knowledge for text generation. Moreover, we vary the number of diffusion steps since it is a trade-off between synthesis quality and efficiency. Surprisingly, increasing the diffusion steps will not lead to performance gains. We speculate that diffusion with certain steps can provide enough visual knowledge for the PLM, and more steps may just help to achieve higher resolution. Thus, we only synthesize for 25 steps considering the efficiency.

5.5 Human Evaluation

Considering that the automatic evaluation may be inconsistent with human judgments, we further invite five college students to assess the generated texts. We randomly choose 100 samples from the test set of each dataset and showcase the generated texts of both BART and LIVE-BART. The annotators should choose which one is better or choose a tie based on their subjective feelings. From the results in Table 6, we can observe that our LIVE method can make BART generate more satisfactory texts in all tasks.

6 Conclusion

In this paper, we present the **LIVE** method for natural language generation. First, we propose an imagination-based method, imitating the process of human writing. It is a relevant, selective, and dynamic approach that leverages Stable Diffusion to synthesize images for each input sentence and discard the images with lower text visuality computed by CLIP. Furthermore, we introduce a plug-and-play vision-text fusion layer to deeply incorporate visual knowledge into PLMs and obtain visually-augmented text representations for text generation.

Extensive experiments have demonstrated that our LIVE methods are compatible with two PLMs (*i.e.,* BART and T5) and can achieve superior performance over all the baseline models.

In future work, we will investigate how to synthesize more relevant images based on the input prompt and design a finer fusion method for better aligning different words and images. We will also attempt to extend our methods to more tasks (*e.g.,* language understanding) and PLMs (*e.g.,* BERT). Besides, it is meaningful to explore the probability of combining our LIVE method with existing large language models (Zhao et al., 2023) to enhance their representation and generation capabilities.

Acknowledgment

This work was partially supported by National Natural Science Foundation of China under Grant No. 62222215, Beijing Natural Science Foundation under Grant No. 4222027, and Beijing Outstanding Young Scientist Program under Grant No. BJJWZYJH012019100020098. Xin Zhao is the corresponding author.

Limitations

We only conduct experiments on four natural language generation tasks without considering the expandability to more NLP tasks, such as language understanding or reasoning. It is also meaningful to investigate the robustness of our methods with different text formats (*e.g.,* text length and literary form), *i.e.,* examine which situations and why our methods can achieve better performance. Due to the limitation of computing power, we do not explore the effectiveness of our methods under different PLMs with various scales. Besides, we utilize CLIP to evaluate the text visuality and encode images into representations, and this is also interesting to research which vision encoder has higher suitability with PLMs.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2383–2395, Dublin, Ireland. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5698, Dublin, Ireland. Association for Computational Linguistics.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang, and Ji-Rong Wen. 2022. Visually-augmented pretrained language models for nlp tasks without images. *arXiv preprint arXiv:2212.07937*.

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 *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. 2021. Probing contextual language models for common ground with visual representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5367–5377, Online. Association for Computational Linguistics.

Anubhav Jangra, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2021. A survey on multi-modal summarization. *arXiv preprint arXiv:2109.05199*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Woojeong Jin, Dong-Ho Lee, Chenguang Zhu, Jay Pujara, and Xiang Ren. 2022. Leveraging visual knowledge in language tasks: An empirical study on intermediate pre-training for cross-modal knowledge transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2750–2762, Dublin, Ireland. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022b. A survey of pretrained language models based text generation. *arXiv preprint arXiv:2201.05273*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, 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 *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A visual experience powered conversational agent. 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 5596–5611, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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 *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yujie Lu, Wanrong Zhu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-augmented natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4392–4402, Seattle, United States. Association for Computational Linguistics.

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 *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. 2021. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

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, Melbourne, Australia. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Lei Shen, Haolan Zhan, Xin Shen, Yonghao Song, and Xiaofang Zhao. 2021. Text is not enough: Integrating visual impressions into open-domain dialogue generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4287–4296, New York, NY, USA. Association for Computing Machinery.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Los Alamitos, CA, USA. IEEE Computer Society.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Tianyi Tang, Junyi Li, Zhipeng Chen, Yiwen Hu, Zhuohao Yu, Wenxun Dai, Wayne Xin Zhao, Jian-yun Nie, and Ji-rong Wen. 2022. TextBox 2.0: A text generation library with pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 435–444, Abu Dhabi, UAE. Association for Computational Linguistics.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. In *Advances in Neural Information Processing Systems*, volume 34, pages 24468–24481. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Los Alamitos, CA, USA. IEEE Computer Society.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2022b. Visually-augmented language modeling. *arXiv preprint arXiv:2205.10178*.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☒ A2. Did you discuss any potential risks of your work?
We utilize existing PLMs and datasets for improving text generation, without creating new models or datasets.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

See below

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section A

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section A

C ☑ Did you run computational experiments?

See below

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3.4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.4 and 4.1.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.4, 4.1.2, and 4.1.4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.