Brush2Prompt: Contextual Prompt Generator for Object Inpainting

Mang Tik Chiu[1,2], Yuqian Zhou[2], Lingzhi Zhang[2], Zhe Lin[2],
Connelly Barnes[2], Sohrab Amirghodsi[2], Eli Shechtman[2], Humphrey Shi[1,3]
[1]UIUC, [2]Adobe, [3]University of Oregon

Abstract

Object inpainting is a task that involves adding objects to real images and seamlessly compositing them. With the recent commercialization of products like Stable Diffusion and Generative Fill, inserting objects into images by using prompts has achieved impressive visual results. In this paper, we propose a prompt suggestion model to simplify the process of prompt input. When the user provides an image and a mask, our model predicts suitable prompts based on the partial contextual information in the masked image, and the shape and location of the mask. Specifically, we introduce a concept-diffusion in the CLIP space that predicts CLIP-text embeddings from a masked image. These diffused embeddings can be directly injected into open-source inpainting models like Stable Diffusion and its variants. Alternatively, they can be decoded into natural language for use in other publicly available applications such as Generative Fill. Our prompt suggestion model demonstrates a balanced accuracy and diversity, showing its capability to be both contextually aware and creatively adaptive.

1. Introduction

In traditional background image inpainting [2], the primary goal is to fill in a masked region using the surrounding background context, thus removing any original objects in that region. These methods usually do not take conditions like text prompts and typically do not introduce new elements into the image. However, with the advent of diffusion-based text-to-image models such as Stable Diffusion [25, 32], DALLE2 [30], and Imagen [33], conditional models have been trained to insert objects to images using explicit text prompts. This is known as text-guided inpainting. These methods encode the input text prompt into latent embeddings, which then guide the image diffusion process through cross-attention. With an appropriately designed text prompt, they can generate highly detailed results seamlessly blended with the background. Their remarkable generation capability gained widespread attention spanning from academic circles to various industries.

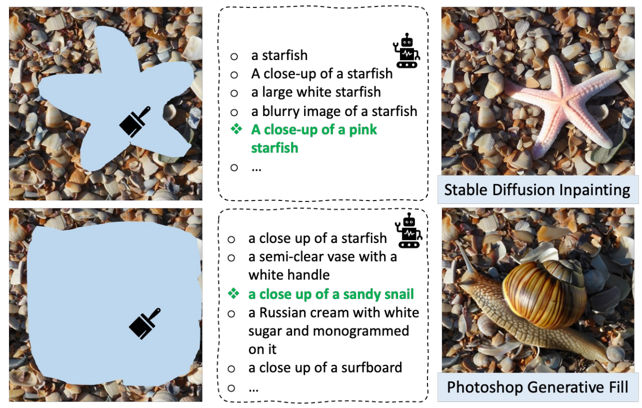

Figure 1. We propose a context-aware prompt generator for text-guided object inpainting task. We provide diverse prompt suggestions by analyzing both the image context and the shape of the mask as soon as users draw a mask on the image. Our generator can be compatible with any text-guided inpainting tools.

Text-guided inpainting can be used for object inpainting, which is the task of adding one or more new objects to an indicated region of an image [42, 47]. In existing models, this requires users to provide explicit text prompts to describe their envisioned concepts. This necessity leads to a question: Can object insertion be achieved without an explicit user text prompt? Object insertion without a text prompt could be beneficial for inexperienced users who might not know what the text-to-image system is capable of, or it might help experienced users generate creative ideas, or it may simply help save time in case a proposed object matches what the user was imagining. However, in existing models, if the user provides an empty text prompt or uses vague terms like "high quality," the model tends to default to sampling dominant contents from its heavily skewed training dataset. For instance, applying a mask to the sky in an image will more likely result in the model filling in with sky textures or clouds, rather than a beautiful bird. One study [3] proposed to sample meaningful diversity using an inference technique that diversifies the outputs by distancing their generation paths from each other.

This approach enables the generation of random objects for inpainting without the need for a text prompt. However, it sacrifices the precision of prompt guidance, resulting in generations that are less controllable by users.

In this paper, we investigate solutions to simplify the process of coming up with prompts for object inpainting. Our proposed Brush2Prompt, a contextual prompt generator, is designed to automatically suggest diverse and meaningful prompts as soon as the user places a mask on the image. We investigate three different kinds of masks that range in how precisely they indicate the shape of an object, including bounding boxes, convex hulls, and tight masks. The primary objectives of our model are centered on three key aspects: *context awareness*, *mask awareness*, and *diverse generation*. *Context awareness* means that the proposed prompts should be plausible given the surrounding context. *Mask awareness* refers to making the prompt match the users' intentions if the mask provides useful shape information. Finally, *diverse generation* is aimed at fostering creativity by generating different object categories and attributes.

To realize context awareness, we developed a multi-modality masked-Image-to-Text (m-I2T) model operating in the CLIP [28] space. This model takes in pretrained CLIP image embeddings of a masked image, and samples appropriate CLIP-text embeddings to be used in the image generator. In order to make our model seamlessly compatible with general text-guided inpainting models, we also trained a text decoder. It translates the embeddings into prompts in natural language form. Additionally, it offers users the flexibility to manually modify these prompts or to utilize an auxiliary auto-completion feature for further ease of use.

To attain mask awareness, we implemented a mask shape augmentation strategy during the training phase. This approach was based on our observation that the shape of a mask can convey significant conceptual information, reflecting the user's intentions. For example, a mask shaped like a car should encourage the model to focus on car-related concepts. Alternatively, a simple bounding box shape should allow a more flexible and creative concept suggestion. The process of generating suggested prompts in our model is stochastic, which facilitates generation of diverse object categories and attributes. To summarize, our contributions of this work are:

- We propose a contextual-aware prompt generator designed for object insertion in image inpainting tasks. It is trained to sample text embeddings given masked images. We employ mask shape augmentation during training to align users' intentions with mask shapes. A prompt decoder is also developed to convert the embedding to natural language prompts. The model is seamlessly compatible with generic text-guided inpainting models, making it a versatile plug-and-play tool.
- We investigate the influential factors of the prompt generation quality: image context and mask shape. Our findings reveal that the accuracy and diversity of the generated results can vary based on different configurations of the inputs and models.
- To evaluate the accuracy and diversity of the generated prompts, we curate and organize the first benchmark dataset Brush2PromptBench. This dataset provides a comprehensive baseline for evaluating the performance of contextual and mask-aware prompt generation in object inpainting.

2. Related work

Diffusion Models. Diffusion models [7, 31, 37] drastically improved the quality of generated images compared to more traditional generative models such as GANs. These models work by learning to reverse an iterative noising process, where random Gaussian noise is added to the original image. As a result, during inference, the trained model can then progressively perform denoising on a randomly sampled Gaussian map and generate images close to the trained data distribution. Following the success of unconditional diffusion models, numerous extensions have been made to enable various use cases. For example, by conditioning the denoising process on encoded text inputs from pretrained vision-language models such as CLIP [28], diffusion models [29, 31, 34] can be used to generate images that correspond to the text description from the user, leading to very impressive results. Furthermore, an additional mask condition can be imposed onto these text-to-image models, where the model is trained to only generate the prompted concept within the masked region. This leads to various text-guided inpainting models [1, 23, 42] that enables even finer control for image generation. Alternatively, [44] proposed to use a reference image instead of a text prompt for more precise style and structure control in the generation process.

Image-to-Text Models. Different from these text-to-image models, researchers also focused on predicting text given an image condition. One such popular task is image captioning, where the goal of the model is to generate an accurate description of objects or the scenery in an image. These models usually consist of an image encoder for feature extraction, and a text generator in the form of RNNs [8, 9, 21], attention-based networks [20, 43, 49], and eventually transformers [11, 12, 14, 41]. More recent approaches directly leverage pretrained vision-language models to extract rich image features, and either train a transformer or fine-tune language models [15, 22, 39, 40] to generate captions. Note that image captioning is fundamentally different from our task, since they focus on describing the scene or subjects in the image, while our task focuses on suggesting reasonable new concepts given a masked image context.

Object Inpainting. Object inpainting [38, 42, 47] shares

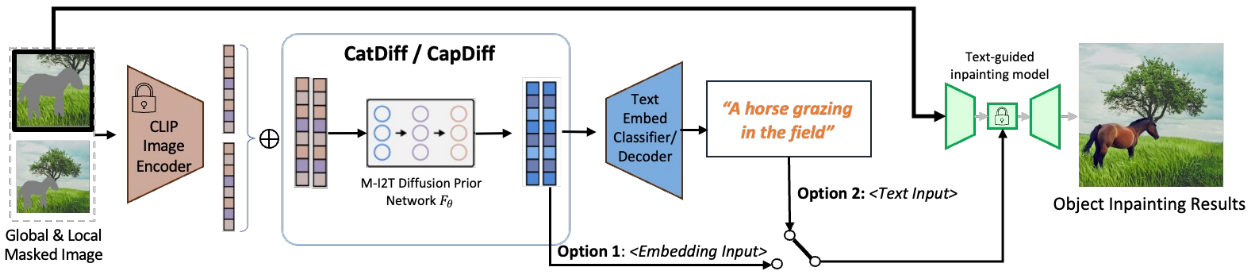

Figure 2. Given an image and mask, we first encode the the global and local crop of the masked image using frozen CLIP image encoder. Then we train the m-I2T diffusion prior network F_θ (CatDiff for category label generation, and CapDiff for caption generation) to generate diverse results aligned with the possible concepts within the masked regions. We further train the embedding classifier or text decoder to translate the embeddings into text prompts. Both the generated embeddings or text can be applied to text-guided image inpainting models.

a similar objective with background inpainting, where the model attempts to fill in missing regions of an image. However, instead of drawing background pixels, object inpainting aims to restore either partially [46] or completely missing [26] objects based on the surrounding context. As opposed to generating objects purely from the input photograph, diffusion models have also been used to composite and harmonize objects extracted from real photos [38]: this is a different problem than we address. Some methods [26, 35] also attempt to construct a scene-graph based on unmasked objects to establish stronger scene correlation for more accurate object prediction, they evaluate their models based on the accuracy of the predicted object to be restored. However, we find that object inpainting, especially in the context of whole-object insertion, is inherently ambiguous, and as a result cannot be properly evaluated with accuracy alone. For example, given a background image of a table and a circular mask, one cannot judge whether the masked object is a coin or an orange. Therefore, we propose to reformulate the problem as a generative problem and encourage creativity of the concept to be generated.

3. Methodology

3.1. Preliminary: Diffusion Models

We follow the same diffusion model formulation as Xie et al. [42]. The diffusion process involves a data sample z_0 (e.g. image, text, or embedding), and an iterative noising/denoising step. In the forward Markov diffusion process $t \in [1..T]$, z_0 is progressively corrupted into z_T by adding noise with a controlled variance $q(z_t|z_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t} z_t, \beta_t \mathbf{I})$, and it gradually approaches a Gaussian $\mathcal{N}(0, \mathbf{I})$. At the same time, z_t can be computed by $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i), \epsilon \in \mathcal{N}(0, \mathbf{I})$. In the backward diffusion process, a model can be trained to either estimate the added noise at each step ϵ_t or directly predict the unnoised sample z_0. During inference, additional samplers [19, 36] can be used to speed up the reverse process.

3.2. Problem Definition

Given an image I and a mask M, our goal is to generate K text prompts $T_{k,k\in\{1..K\}}$ that describe reasonable concepts to be inserted into the image in the masked region. Previous works [26, 35] defined this problem as a deterministic process, and only predicted a single object category. In our work, we reformulate this problem as a stochastic caption generation process. The generated text allows us to leverage powerful pretrained inpainting models for their image generation capability, and at the same time add user interaction in the process by allowing users to modify the predicted descriptions and refine the output image to their desire. Our pipeline design is similar to image-to-text diffusion methods [45], but our goal is to generate novel object descriptions rather than describing existing image context.

3.3. Masked-Image-to-Text (m-I2T) Diffusion Prior

We show our overall pipeline in Figure 2. Given a training sample in a 3-tuple (Image I, Object mask M, Object description T), we first create a masked image $I_M = I \odot M$. We then use a frozen CLIP-image encoder to extract visual features from I_M. Here, we do not take the $<cls>$ token from CLIP-image embeddings, but instead use the 256-D patch embeddings to preserve spatial and local context information. The obtained masked-image embedding e_{I_M} (we use e_M as a shorthand from now on) is used as input condition for the diffusion prior network F_θ.

Inspired by DALLE-2 [29], where they train a diffusion prior network to translate input text embeddings into image embeddings for better alignment in the image space, here we take the opposite direction, and train a diffusion prior that learns to translate the masked-image embeddings e_M into CLIP-text embeddings e_T. The generated embeddings should encode a description of the potential object candidate for object insertion. By leveraging the generative power of the diffusion prior, we can obtain a diverse set of object descriptions for each image. Since the objective of the model is to predict CLIP-text embeddings, it is naturally compatible with all text-guided inpainting models that use CLIP as the prompt encoder, such as Stable Diffusion-v1 [31] and Smartbrush [42]. Next, we describe how to decode the text embeddings into either *simple shorter category labels* or *longer captions* for evaluation.

Category Diffusion and Decoding (CatDiff). In our approach, where the diffusion prior network generates text prompts within the embedding space, it is essential to find a method to decode these prompts into natural language format which can be evaluated independently of any specific text-to-image model. To simplify this process, we initially reframe the task as a category generation problem. In this scenario, the model is trained specifically to diffuse embeddings related to object categories. Following this, the decoder functions as a straightforward category label classifier. Its role is to predict class labels from the generated embeddings. We found that the classifier can be trained to exhibit high accuracy, so we can leverage this classifier as a reference model to evaluate the diffusion prior network.

To train the diffusion prior network for category generation, we can simply encode the class label as text using the CLIP-text encoder. Following [28], we use the prompt template "A photo of a <category>". We define the encoded category embedding as x_0^c, where c stands for "category", and add noise at each step to generate a noised embedding $x_t^c = \sqrt{\bar{\alpha}_t} x_0^c + \sqrt{1-\bar{\alpha}_t}\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The prior network θ then estimates x_0^c conditioned on the noised embedding, timestep and masked-image embedding. We use MSE to compute the loss for the prior network:

$$\mathcal{L}_{category} = \mathbb{E}\big[||x_0^c - F_\theta(x_t^c, t, e_M)||_2^2\big] \tag{1}$$

We then train the classifier to predict object categories conditioned on text embeddings. We use a single transformer block with 2 attention heads and 64 hidden dimensions for the classifier, and a linear layer to map to the categories. During training, the classifier takes in ground truth text embeddings of the masked object, and predicts its category. To train both modules, we can use popular object segmentation datasets such as COCO [17] and OpenImages [10]. The human-labelled segmentation mask and ground truth label pairs are accurate enough for training and evaluation.

Caption Diffusion and Decoding (CapDiff). To further improve the diversity of the results and make the model more practical in real user cases, we extend the work to diffusion for the caption and translate the embeddings using text decoders. The diffusion prior network can be trained in a similar way as category diffusion, where we simply replace category labels with longer descriptive local captions:

$$\mathcal{L}_{caption} = \mathbb{E}\big[||x_0^d - F_\theta(x_t^d, t, e_M)||_2^2\big], \tag{2}$$

where d stands for "description".

Decoding the caption embeddings is more complicated than embedding classification. In our experiments, we find that directly predicting text tokens from each CLIP-text embedding leads to poor quality text outputs. As a result, we opt to instead finetune a pretrained text decoder to translate the embeddings into a caption.

Specifically, we finetune a pretrained GPT-2 [27] text decoder that takes in the generated clip-text embedding as prefix. The model is then trained to predict a caption which conveys the object encoded in the embeddings. Similar to the category diffusion set up, we train the text decoder using the ground truth caption and the corresponding CLIP-text embeddings. During inference, we can then feed the generated text embedding from the diffusion prior and obtain the corresponding generated caption. Following [22], we train the text decoder using cross entropy loss, where the model tries to predict the next text token given the CLIP-text embedding as prefix and previous text tokens. The objective of text decoder is:

$$\max_\phi \sum_{i=1}^{L} \log p_\phi(w_i | x_0^d, w_0, w_1, ..., w_{i-1}), \tag{3}$$

where ϕ is the text decoder parameters, L is the token length of the sentence, and $w_i, i \in [1..L]$ are the text tokens of the caption.

3.4. Context and Diversity

Context Control via Global-Local Image Conditions In our experiments, we found that the size of the mask relative to the image can also impact the final concept generation quality. If we use the entire global image as input and the masked area is too small or off-centered, the model tends to ignore the mask shape and region, and instead generates concepts related to the global context or other objects in the image. On the other hand, if we crop tightly around the hole regions, the input of the model will lack in global context and generate some concepts irrelevant to the original image. Therefore, we propose to use both the global and local CLIP image embedding by concatenating them as the inputs to our diffusion model. This approach achieves the balance between global context and shape precision, yielding overall better accuracy and diversity.

Diversity Control via Mask Shape Augmentation The shape of the input mask can sometimes provide strong hints to certain object categories. For example, a mask that closely resembles an object category (e.g. elephant) should provide a stronger constraint on the output variety of the model, while a simple bounding box should allow for higher diversity. To enable control of the concept diversity, we randomly augment the shape of the mask during training.

Specifically, for each training sample, we randomly augment the mask into one of four shapes, which are the original tight mask for precise control, a dilated mask for approximate control, a dilated convex hull for loose hints, and a bounding box for maximum diversity. We then separately evaluate them to study the impact of the mask shape towards concept diversity.

4. Experiments

4.1. Implementation Details

Datasets. As mentioned in Section 3, we use the COCO instance segmentation dataset [17] for classification and a subset of the OpenImages [10] dataset for both classification and caption generation. COCO contains around 118K/5K training/validation images and 80 categories with instance segmentation masks. Our collected OpenImages subset contains around 935K/13.4K training/testing images, where all images selected are larger then 512×512. Training the diffusion prior network for object caption generation requires local captions that describe the masked object. We follow [42] and use BLIP [13] to obtain local object captions from the OpenImages dataset. Since the local captions are not always accurate, we additionally apply a strict filter that removes all object masks where the caption does not contain the label category. The resulting mask-caption pair then allows us to train and evaluate the models. We have named our testing partition *Brush2PromptBench* and plan to release this test set to the research community.

Model Details. We use a transformer-based architecture for the prior network. The transformer has 12 attention blocks, each with 12 attention heads and 128 hidden dimensions. Both input and output embeddings have a dimension of 768, which is the same as CLIP and can therefore be directly injected into text-guided inpainting models with CLIP embeddings as the text inputs. Similar to [29], we train the model to directly predict the noise-free embedding. For sampling at inference time, we use 10 and 50 iterations with DDIM sampler [37] for CatDiff and CapDiff respectively. Both CatDiff and CapDiff models, including the respective decoders, are trained for 20K iterations with an effective batch size of 1024 and a learning rate of 10^{-4}. We train the diffusion prior and the decoders separately.

During training, we randomly crop 512×512 patches from each image, then select a random instance mask within the cropped region. We use the 512×512 region to calculate the global CLIP embedding and then center crop around the masked region with $1.5\times$ expansion (e.g. if mask has a size of 100×100, we center crop 150×150 and resize to 512×512) around the hole to form the local CLIP embedding. To avoid masks that are too small or too large, such that the context is either unrelated or completely lost, we further filter masks that are smaller than 1% or larger than 50% of the image area. During testing, center crop is used to obtain 512×512 images. Unless otherwise specified, the first instance in each image by index is used for evaluation.

Baselines. As far as we are aware, our diverse prompt recommendation pipeline is novel in this field, distinct from prior research works. Some related studies like [26, 35] were conducted under more constrained conditions. For instance, they might focus on a limited subset of class categories and images, and their evaluations are primarily centered on classification accuracy without considering the diversity of concepts. Therefore, we compare our task-specific model with recent generic visual-language instruction tuning models, including **BLIP-VQA** [13] [1], **InstructBLIP** [4] [2] and **LLaVA** [18] [3]. We treat these multi-modal instruction models as generic language agents that are both aware of image context and have high domain-diversity, thus they are the most suitable candidates to compare with in this novel task compared to models trained on limited domains, such as image captioning models. Prompting these models can be tricky, as it requires careful adjustment of the questions to achieve the desired results. For instance, we might pose a question such as "Write one text prompt that describes reasonable objects to be inserted in the gray area." This approach is used to guide LLaVA to either return one prompt each time, a method we refer to as **LLaVA-Resample**, or to respond with five prompts at once, which we call **LLaVA-5-Prompt**, aligning more closely with our experiments. Detailed explanations and methodologies related to these prompting strategies are available in the supplementary material.

Metrics. For CatDiff experiments, we use K-1 and K-5 classification accuracy to measure *context awareness*. K-n means we sample n predictions and see whether one of them match with the ground truth category of the masked region. For *generation diversity*, we propose to use K-50 entropy, which means we sample 50 predictions, and compute the average entropy of the predicted class probabilities. For CapDiff, we follow [6] and report BLEU [24], ROUGE [16], BERTScore [48] of the generated sentences to evaluation *context awareness*. We also report Dist-1 [6], Self-BLEU [50] and Div-4 [5] to evaluate caption diversity. For each image, we sample five captions for evaluation. The evaluation is also conducted on different levels of mask coarseness to validate *mask awareness*.

4.2. Quantitative and Qualitative Results

CatDiff Results. The class category prediction results for our model on the COCO and OpenImages datasets are presented in Figure 3. We evaluated the model using three different mask types: tight mask, convex hull, and bounding

[1] https://github.com/salesforce/LAVIS#visual-question-answering-vqa
[2] https://huggingface.co/docs/transformers/model_doc/instructblip
[3] https://github.com/haotian-liu/LLaVA

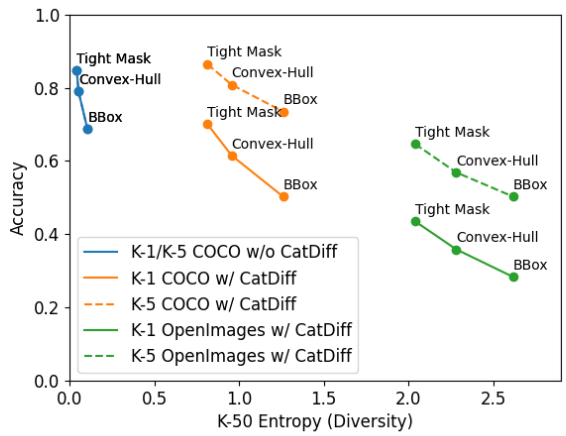

Figure 3. Classification results on COCO and OpenImages. Our diffusion prior network CatDiff has significantly higher diversity compared to a deterministic network, and achieves similar accuracy when sampled for 5 times. As expected, the diversity of category generation increases as we relax the mask shape constraint.

box. Our findings indicate a decrease in prediction accuracy as the mask becomes coarser. However, the entropy, which represents the diversity of the results, increases with coarser masks. This trend can be attributed to the fact that larger masks reduce the amount of contextual information available from the image, as well as the conceptual information provided by the shape of the mask. As a result, when the mask is enlarged, the CatDiff model compensates by generating a broader range of diverse results. This balance between accuracy and diversity is a key aspect of our model's performance, demonstrating its adaptability to varying levels of contextual and conceptual input.

CapDiff Results. In Table 1 and Figure 4, we present a comparison of our model with open-sourced generic visual-language models. It is important to note that prompting these models in various ways can yield different outcomes, and fine-tuning the questions to achieve optimal results can be challenging. Under our specific testing settings and the questions we formulated, baseline models—which are typically trained for general visual-language tasks—tend to score lower on our benchmark dataset. This lower performance could be attributed to their training, which focuses on extracting information from complete image contexts and describing existing objects, rather than understanding partial images tailored to our specific task. This highlights both the uniqueness and the challenges inherent in our proposed task. In contrast to these baselines, our model demonstrates superior performance, achieving higher scores in BLEU, ROUGE, and BERTScore metrics. These results suggest that the captions generated by our model are more closely aligned with the original masked objects. This alignment indicates that our method is more likely to meet users' intentions. Meanwhile, our model's performance in metrics that measure sentence-level diversity is either higher or comparable to the baselines. This is a significant observation, as it underscores our model's ability to generate a variety of different sentence structures and ideas. However, the word diversity in our model's outputs (i.e. Dist-1) is not at an optimal level. This limitation could be attributed to the size of our training dataset. Currently, our dataset might not provide a sufficiently wide range of vocabulary and concepts to enhance word-level diversity. This could be improved by expanding the training prompts. We also demonstrate the full pipeline results in Figure 5.

4.3. Ablation Studies

Diffusion Prior We assert that incorporating a diffusion prior into our prompt suggestion pipeline, which includes both CatDiff and CapDiff models, significantly enhances the diversity of prompt suggestions for the masked area in an image, particularly when the context image is processed through a CLIP-image embedding of the masked image. To validate this hypothesis, we conducted an ablation study by omitting the diffusion prior from the process.

For category prediction, we trained a deterministic transformer that shares the same architecture as the prior network to predict category embeddings without diffusion steps. For caption prediction, we finetuned a pretrained GPT-2 decoder to generate captions directly from the CLIP embeddings of the masked image. To introduce variability in the outputs, we utilized multinomial sampling without beam search during the inference phase. The results are shown in Table 3 and 4.

As shown in Table 3 and Figure 3, a deterministic transformer without CatDiff demonstrates a high level of accuracy in predicting the actual category behind the mask. However, it significantly lacks diversity in its outputs. This limitation is critical since it does not suggest alternative possibilities that could also be contextually appropriate. In contrast, models with CatDiff sample various reasonable categories that can be inserted into the region, and even attain slightly higher accuracy when multiple samples are drawn. The results for caption generation also highlight the effectiveness of CapDiff in enhancing the diversity of generated captions. A few examples are illustrated in Figure 4. More are in the supplementary material.

Global-Local Embeddings. In Section 3.4 of our paper, we mentioned the importance of integrating both global and local contexts to enhance the accuracy of context comprehension in our model. We validate this in experiments on CatDiff. As depicted in Table 2, the use of combined global-local embeddings leads to an overall improvement in accuracy, particularly in the K-1 accuracy metric. This finding underscores the effectiveness of our approach in accurately

Mask shape	Method	BLEU ↑	ROUGE ↑	BERTScore ↑	Dist-1 ↑	Self-BLEU ↓	Div-4 ↑
Tight mask	BLIP-VQA [13]	0.005	0.071	0.537	0.998	0.600	0.038
	InstructBLIP [4]	0.071	0.308	0.704	0.882	0.466	0.570
	LLaVA-Resample [18]	0.031	0.275	0.684	0.955	0.286	0.715
	LLaVA-5-Prompt [18]	0.023	0.269	0.656	0.998	0.163	0.785
	CapDiff (Ours)	0.177	0.427	0.732	0.845	0.169	0.858
BBox	BLIP-VQA [13]	0.004	0.053	0.526	0.999	0.627	0.025
	InstructBLIP [4]	0.060	0.280	0.691	0.881	0.438	0.597
	LLaVA-Resample [18]	0.026	0.254	0.678	0.970	0.312	0.694
	LLaVA-5-Prompt [18]	0.020	0.245	0.648	0.998	0.128	0.881
	CapDiff (Ours)	0.149	0.383	0.715	0.838	0.141	0.885

Table 1. Caption generation results on the OpenImages dataset. BLEU [24] and ROUGE [16] evaluates text quality. BERTScore [48] further evaluates the semantics between the predicted and ground truth sentences. Following [6], Dist-1 measures the word diversity within a sentence, whereas Self-BLEU [50] and Div-4 [5] measure the sentence level diversity of a group of sentences. For each experiment, we generate five candidate text prompts for evaluation. For LLaVA [18], we either ask the same questions five times (LLaVA-Resample) or instruct it to generate five prompts directly (LLaVA-5-Prompt). Top ranked result is indicated in blue; second best result indicated in green.

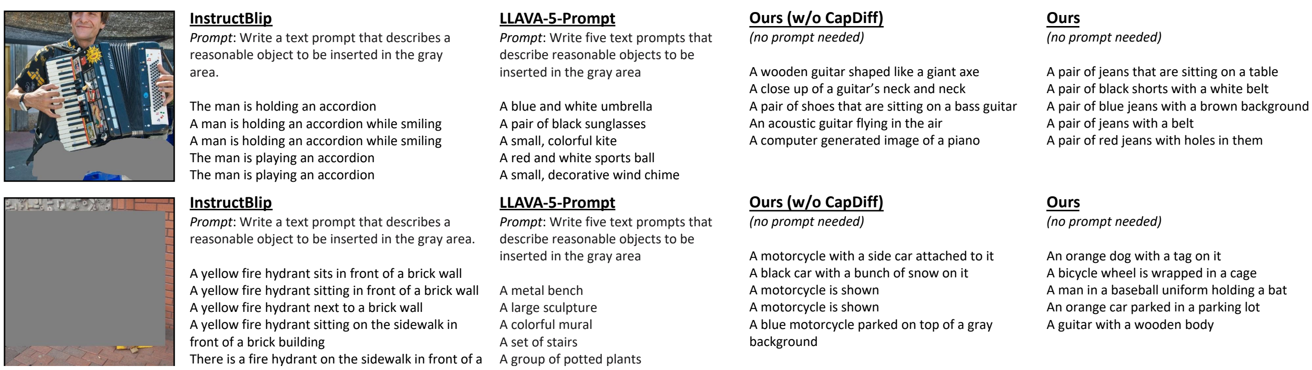

Figure 4. Qualitative results. Compared to InstructBlip [4], LLaVA [18], and a baseline without CapDiff, our approach generates prompts that are diverse and context-aware. We show the five generated prompts for each model.

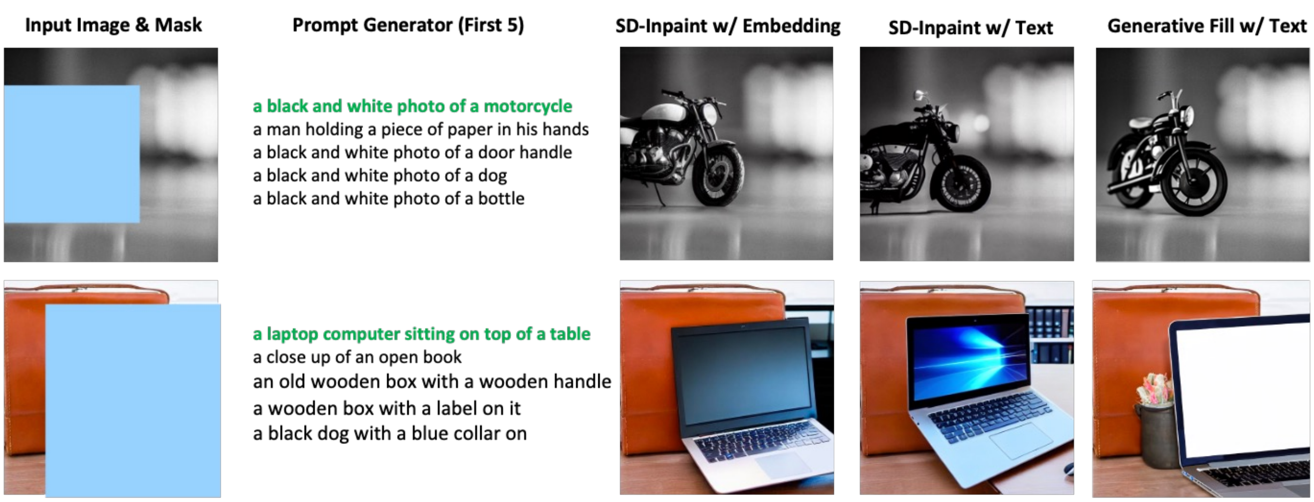

Figure 5. Full pipeline demonstration. Given an image and mask, we propose multiple diverse prompt suggestions. Users can select one of them, and apply it to any text-guided inpainting tool, either with embedding or the decoded text prompts.

Mask size	Embed type	K-1 Accuracy	K-5 Accuracy	K-50 Entropy
Regular	Global	0.681	0.847	0.884
	Local	0.681	0.843	0.845
	Global-Local	0.700	0.864	0.815
Small	Global	0.604	0.804	1.073
	Local	0.613	0.799	1.027
	Global-Local	0.692	0.800	0.490

Table 2. Ablation on input embedding types on COCO classification dataset. Top result is in blue; second result is in green.

identifying the most relevant categories for a given context. Additionally, Figure 6 depicts up to 5 top categories after predicting 50 samples. This demonstrates a decrease in diversity, which can be attributed to the model's improved capability to eliminate unrelated concepts. This improvement occurs when there is limited context available (as in the case of local embeddings) or when the mask size is relatively small (as in global context scenarios). The reduction in diversity, in this case, is not a drawback but rather an indication of the model's ability to focus on relevant concepts and disregard those that are less pertinent to the given context. This balance between accuracy and diversity is a key aspect of the model's performance, demonstrating its capability to adapt to varying levels of contextual information.

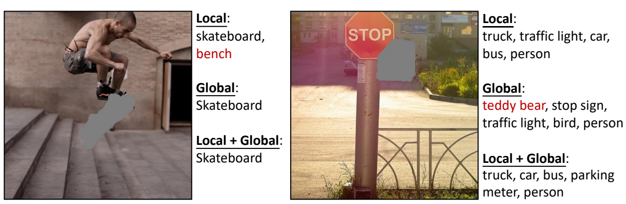

Figure 6. Qualitative comparison of diffusion categories under different contextual conditions, with red marking less-relevant suggestions. The top 5 (if any) unique categories after predicting 50 samples are shown. Integrating both local and global contexts usually better balances contextual relevance with diversity.

5. Extensions and Limitations

Our models can be extended to context-aware prompt completion tasks. However, in scenarios where users already provide initial prompts, the CapDiff component might not be essential. To address this, we developed a method where the CLIP-image embedding from the masked image is directly injected into a GPT-2 decoder. This approach is designed to efficiently complete prompts based on the given context and initial user input. The details of the experiments and results are shown in the supplementary material. Additionally, our models currently have limitations in terms

Mask shape	CatDiff	K-1 Accuracy	K-5 Accuracy	K-50 Entropy
Tight Mask	✗	0.847	0.847	0.043
	✓	0.700	0.864	0.815
Convex Hull	✗	0.790	0.790	0.053
	✓	0.615	0.808	0.957
Bounding Box	✗	0.687	0.687	0.106
	✓	0.503	0.735	1.260

Table 3. Ablation on CatDiff using COCO. Our diffusion prior network leads to higher diversity compared to a deterministic network, and achieves similar accuracy when sampled more.

Mask shape	CapDiff	BLEU ↑	BERT-Score ↑	Self-BLEU ↓	Div-4 ↑
Tight mask	✗	0.249	0.785	0.215	0.808
	✓	0.177	0.732	0.169	0.858
Bounding box	✗	0.193	0.758	0.167	0.855
	✓	0.149	0.715	0.141	0.885

Table 4. Ablation on CapDiff. Diffusion prior network leads to higher diversity. On the other hand, it also results in deviation from the original caption, which leads to lower alignment scores.

of vocabulary depth. This aspect could be improved by incorporating more diverse data into the training process. Looking ahead, we are interested in exploring more complex multi-modal visual-language architectures which have the potential to significantly enhance the quality of generation, making the models more robust and versatile.

6. Conclusion

In our paper, we introduced a novel task on generating meaningful and diverse prompts for object inpainting. We identified three critical aspects for evaluating our model: *context awareness*, *shape awareness*, and *diverse generation*. To effectively incorporate image contextual information while also enhancing the diversity of prompt generation, we employed diffusion prior modules—CatDiff and CapDiff—on top of CLIP image embeddings of masked images. Our experiments demonstrated the effectiveness of this approach through accurate and diverse category label and caption generation. We developed a classifier for category generation and a text decoder for caption generation. These components not only aid in the inspection and evaluation of results but also make our generator a plug-and-play tool. Our research showed that our task-specific model surpasses generic visual-language models in caption generation. Looking forward, we see potential in applying these ideas to more complex architectures and expanding the training datasets.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 1

[3] Noa Cohen, Hila Manor, Yuval Bahat, and Tomer Michaeli. From posterior sampling to meaningful diversity in image restoration. *arXiv preprint arXiv:2310.16047*, 2023. 1

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 5, 7

[5] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10695–10704, 2019. 5, 7

[6] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. 5, 7

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[8] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision*, pages 2407–2415, 2015. 2

[9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[10] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2016. 4, 5

[11] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 2

[12] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8928–8937, 2019. 2

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5, 7

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[15] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023. 2

[16] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5, 7

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4, 5

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 5, 7

[19] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 3

[20] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 2

[21] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE international conference on computer vision*, pages 2533–2541, 2015. 2

[22] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 4

[23] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5, 7

[25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[26] Jiayan Qiu, Yiding Yang, Xinchao Wang, and Dacheng Tao. Hallucinating visual instances in total absentia. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 264–282. Springer, 2020. 3, 5

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 3, 5

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 2, 4

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[35] Tripti Shukla, Paridhi Maheshwari, Rajhans Singh, Ankita Shukla, Kuldeep Kulkarni, and Pavan Turaga. Scene graph driven text-prompt generation for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 759–768, 2023. 3, 5

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[38] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 2, 3

[39] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 2

[40] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022. 2

[41] Yiyu Wang, Jungang Xu, and Yingfei Sun. End-to-end transformer based model for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2585–2594, 2022. 2

[42] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 1, 2, 3, 4, 5

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[44] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023. 2

[45] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 3

[46] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 3

[47] Yu Zeng, Zhe Lin, and Vishal M Patel. Shape-guided object inpainting. *arXiv preprint arXiv:2204.07845*, 2022. 1, 2

[48] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 5, 7

[49] Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J Corso. Watch what you just said: Image captioning with text-conditional attention. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 305–313, 2017. 2

[50] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018. 5, 7