A. Discussion about NOIC and ZIC

In this section, we will discuss the difference between novel object image captioning (NOIC) and zero-shot image captioning (ZIC), brief comparisons are shown in Table 1 and details are as follows:

- **Generalization among objects vs. among tasks.** NOIC aims to generalize image captioning (IC) models to “novel objects” not presented in the training images. This means both training and testing tasks are IC. By contrast, the “zero-shot” concept in our work (and most related work in paper) comes from GPT-3, referring to applying large pre-trained models for downstream IC tasks with no task-specific fine-tuning.

- **With vs. without curated training image-caption pairs.** NOIC models are often trained on well-designed image-caption pairs of seen objects. Hence, different dataset splits are often considered to perform evaluation. By contrast, ConZIC is free of well-designed image-caption pairs to perform training or even fine-tuning.

- **With vs. without extra knowledge.** NOIC methods often learn the relations between objects and extra tags, such as attributes and class embeddings. Then, these relations are generalized to unseen objects by various techniques. By contrast, ConZIC utilizes the knowledge from large pre-trained models and thus is free of extra information.

B. Algorithm of Gibbs-BERT

After randomly choosing the generation order, Gibbs-BERT starts from a full noisy sentence (e.g., all [MASK] tokens). At each iteration, Gibbs-BERT progressively samples each word by putting [MASK] at this position and then selecting the top-1 word from the predicted word distribution over the vocabulary by BERT. The result of t-th iteration is the initialization of the (t + 1)-th iteration. The pseudo-code is shown in algorithm 1.

C. SketchyCOCO caption benchmark

SketchyCOCO caption is a small sketch-style image captioning benchmark based on SketchyCOCO, including 14 classes, as shown in Fig. 3. SketchyCOCO is not an image captioning dataset since it only has the classification label. We construct the captioning benchmark through the following steps: i) randomly sample 100 sketch images for each foreground class. ii) label them with a simple prompt, i.e. “A drawing of a [CLASS]”, where [CLASS] is the class name. For example, a cat image is labeled as “A drawing of a cat.”. More details can be seen in Appendix C.
oil paintings and cartoon images. Besides, our method is proven to have efficient application in images with abundant world knowledge, e.g. medical, geography, celebrity, and artworks.

E. Controllable tasks

E.1. Diversity of length control

Table 2 has reported diversity performance where we select the best-1 caption on each length and then compute diversity metrics on these four lengths. Our method surpasses ZeroCap by a large margin.

E.2. Infilling tasks

We have conducted experiments on one-word-infilling and multiple-word-infilling.

**One-word-infilling.** We randomly corrupt one verb/noun in the reference caption, and ask models to infill the most suitable word given other words. We use three metrics to evaluate the accuracy performance: 1) BLEU-1(B-1) to measure unigram precision; 2) Wordnet path similarity(WSim) which measures node distance in Wordnet. Especially, this metric can only be computed between two words of the same POS. Therefore, we set WSim as 0 when the answer has a different POS from the reference word; 3) BERT word similarity(BSim). We use cosine distance in BERT word embedding space, where words have similar semantics generally possess a low distance. Due to its autoregressive nature, ZeroCap can only take the left context into account, which limits its performance. Qualitative results of one-word-infilling are shown in Fig. 1.

**Multiple-word-infilling** In contrast to one-word-infilling, we try to corrupt more words in reference caption. Results with different corrupted ratios are shown in Table 3. We can see that results of a higher corrupted ratio are generally higher in CLIP-S and lower in other metrics.

E.3. Humorous-Romantic control on FlickStyle10k

Quantitative results are shown in Table 5. As we can see, our method has comparable performance in producing captions in specific styles, i.e. romantic and humorous, as shown in Acc column.

E.4. Parts-of-speech controlling

we have tried another POS sequence, ADP DET ADJ/NOUN NOUN NOUN DET ADJ/NOUN NOUN VERB ADV
### Figure 3. Examples of SketchyCOCO caption benchmark

- A drawing of a dog.
- A drawing of a horse.
- A drawing of a cat.
- A drawing of a sheep.
- A drawing of a cow.
- A drawing of a bicycle.
- A drawing of a car.
- A drawing of a motorcycle.
- A drawing of an airplane.
- A drawing of a traffic light.
- A drawing of a fire hydrant.
- A drawing of an elephant.
- A drawing of a zebra.
- A drawing of a giraffe.

### Figure 4. Comparison with groundtruth on MSCOCO caption.

<table>
<thead>
<tr>
<th>GT:</th>
<th>Ours:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Several clocks display the time in different time zones.</td>
<td>Watching some magical silver wall clock highlighting the different time zones within a museum.</td>
</tr>
<tr>
<td>A train sitting in a train station on top of railroad track.</td>
<td>A prototype 1960s general electric Scottish locomotive maintained at the Victoria train shed.</td>
</tr>
<tr>
<td>A white bird walking through a shallow area of water.</td>
<td>A spotted bird shown on a sandy platform with wavy reflections.</td>
</tr>
</tbody>
</table>

### Table 5. Stylized image captioning (i.e., romantic, humorous) performance comparisons on the Flickstyle8k dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Romantic</th>
<th></th>
<th>Humorous</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B-3↑</td>
<td>M↑</td>
<td>CLIP-S↑</td>
<td>Acc↑</td>
</tr>
<tr>
<td>StyleNet</td>
<td>1.5</td>
<td>4.5</td>
<td>-</td>
<td>37.8</td>
</tr>
<tr>
<td>MSCap</td>
<td>2.0</td>
<td>5.4</td>
<td>-</td>
<td>88.7</td>
</tr>
<tr>
<td>MemCap</td>
<td>4.0</td>
<td>7.7</td>
<td>-</td>
<td>91.7</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>1.2</strong></td>
<td><strong>6.1</strong></td>
<td><strong>1.02</strong></td>
<td>96.3</td>
</tr>
</tbody>
</table>
**ZeroCap**(beam 5):
A dog replica.
A dog sculpture.
A dog statue.
A dog sculpture created in London’s Museum of Modern Art.
A dog sculpture created in London’s Museum of Modern Art in the early 2000s.

**Oursshuffle**:
Order: 7, 3, 2, 8, 5, 6, 9, 4, 0, 1
Cap: A striped 3d pet model-sized grey lab tiger displayed.
Order: 7, 8, 1, 5, 3, 4, 2, 0, 9, 6
Cap: A grey metallic 3d model exhibiting a striped pet tiger.
Order: 6, 8, 9, 7, 5, 3, 0, 4, 1, 2
Cap: A tiger sculpture painted on a statue display shown throughout campus.
Order: 5, 9, 3, 4, 6, 7, 2, 8, 1, 0
Cap: A silver painted animal in striped yellow within window displays.
Order: 1, 5, 6, 0, 9, 4, 7, 2, 8, 3
Cap: A silver striped tiger model depicted on window shopping display.

Figure 5. Diversity results compared with ZeroCap.

**VERB ADV**. Results are shown in Table 4.

**F. Bad case analysis**

As shown in Fig. 5, ZeroCap and our method both ignore the “scissor” around the “tiger statue”, which means that how to control which image content to be described, in particular, small objects, is under-explored for zero-shot image captioning.

Besides, as shown in Fig. 2, ConZIC can produce diverse captions in different generation orders, but in some cases, the generation results can not be satisfactory.
Figure 7. Results of images containing world knowledge.