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.*** NOC 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 our paper) comes from GPT-3, referring to *applying large pre-trained models* (trained with no specific task) *for downstream IC tasks with no task-specific fine-tuning*.

• ***With vs. without curated training image-caption pairs.*** NOC 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.*** NOC methods often learn the relations between objects and extra taggers, 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.

Algorithm 1: Algorithm of Gibbs-BERT.

Data: initial sentence: $\boldsymbol{x}^0_{<1,n>} = (x^0_1, ..., x^0_n)$;
iterations=T, candidates=K;
position sequence $P = \textit{Shuffle}([1, ..., n])$;
Result: the final sentence: $\boldsymbol{x}^T_{<1,n>} = (x^T_1, ..., x^T_n)$;
for *iteration* $t \in [1, ..., T]$ **do**
  state: $\boldsymbol{x}^{t-1}_{<1,n>} = (x^{t-1}_1, ..., x^{t-1}_n)$;
  for *position* $i \in P$ **do**
    1. Replace x^{t-1}_i with [*MASK*];
    2. Predict the word distribution over vocabulary by BERT: $p(x_i|\boldsymbol{x}^{t-1}_{-i})$;
    3. Sample x_i from distribution $p(x_i|\boldsymbol{x}^{t-1}_{-i})$;
    4. Replace x^{t-1}_i with x^t_i;
  end
  state: $\boldsymbol{x}^t_{<1,n>} = (x^t_1, ..., x^t_n)$;
end

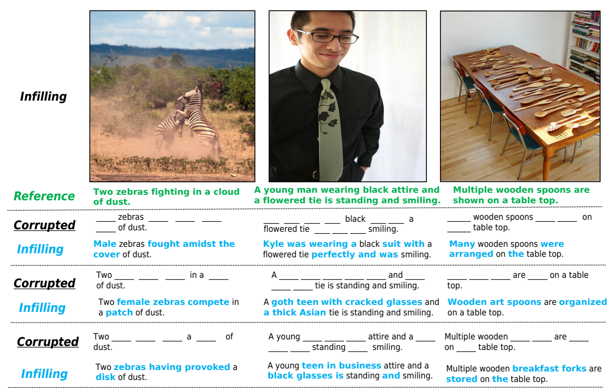


Figure 1. Examples of **infilling task** by ConZIC. Given an image with a reference, we randomly corrupt some words. ConZIC infills these blanks to generate reasonable descriptions, where the infilling words are highlighted in blue.

D. More generation examples

Comparison with Ground-Truth. As shown in Fig. 4, due to the zero-shot nature, caption generations of our method are different from MSCOCO ground-truth. our method has shown significant differences with ground-truth in syntactic(sentence patterns) and semantic(diverse words).

Diverse generation compared with ZeroCap. Comparison results on diverse caption generation are shown in Fig. 5. ZeroCap generates diverse captions by beam search, which can result in a similar sentence pattern with respect to mode collapse. In contrast, our method can produce multiple captions related to the same image by shuffling the word generation order, which has shown strong performance in syntactic and semantic diversity.

Results on various image styles and world knowledge. As shown in Fig. 6 and Fig. 7. Our method performs well in various image styles, *e.g.* natural images, medical images,

	Novel object captioning	ZeroCap or ConZIC
generalization ability	seen objects IC $\rightarrow$ unseen objects IC (with limited background/image styles)	large pretrained models $\rightarrow$ image captioning task (with no limitations on objects/background/image styles)
well designed image-caption pairs	needed for training or fine-tuning	no need
extra knowledge	object taggers	no need

Table 1. Comparisons on problem scenarios of NOIC and our ZIC.

Diversity Metrics	S-C(↑)	Div-1(↑)	Div-2(↑)
Zerocap	0.63	0.40	0.56
Ours	**0.95**	**0.63**	**0.84**

Table 2. Length controlled diversity metrics of our method on MSCOCO. we select the best-1 caption on each length and then compute diversity metrics conditioning on these four captions.

Corrupted Ratio	B-4(↑)	M(↑)	CLIP-S(↑)
0.25	60.69	44.99	0.83
0.50	26.08	29.12	0.89
0.75	8.06	17.60	0.93

Table 3. Results of multiple-words-infilling task.

Parts-of-speech	M(↑)	C(↑)	CLIP-S(↑)	Acc(↑)
without POS	11.54	12.84	1.01	15.54
with POS	7.99	9.29	0.95	86.20

Table 4. Results of parts-of-speech control on MSCOCO. Predefined POS tags is *ADP DET ADJ/NOUN NOUN NOUN DET ADJ/NOUN NOUN VERB VERB ADV*

Figure 2. Example of bad case

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*

Figure 3. Examples of SketchyCOCO caption benchmark

GT:
Several clocks display the time in different time zones.

Ours:
Watching some magical silver wall clock highlighting the different time zones within a museum.

GT:
A train sitting in a train station on top of railroad track.

Ours:
A prototype 1960s general electric Scottish locomotive maintained at the Victoria train shed.

GT:
A white bird walking through a shallow area of water.

Ours:
A spotted bird shown on a sandy platform with wavy reflections.

Figure 4. Comparision with groundtruth on MSCOCO caption.

	Romantic				Humorous			
Methods	B-3(↑)	M(↑)	CLIP-S(↑)	Acc(↑)	B-3(↑)	M(↑)	CLIP-S(↑)	Acc(↑)
StyleNet	1.5	4.5	-	37.8	0.90	4.30	-	41.9
MSCap	2.0	5.4	-	88.7	1.90	5.30	-	91.3
MemCap	**4.0**	**7.7**	-	91.7	**4.0**	**7.20**	-	**97.1**
Ours	1.2	6.1	**1.02**	**96.3**	1.2	6.1	**1.02**	91.4

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

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.

Ours(shuffle) :
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.

GRIT: A series of clocks on top of a screen
CLIPCap: A picture of a skull and crossbones with a lot of wires.
ViTCap:a close up of a cake with a picture of a dog.
ZeroCap: A patient submitted to the neurologist's office.
Ours:
A complete CT medical photo with brain samples.
A CT scene multiple frames displaying a diagnosis.
A stacked sheet displaying signs of brain damaged.
A composite present multiple images featuring frontal trauma.

GRIT: A painting of a painting with a tree in the background
CLIPCap: The night sky over the city.
ViTCap:A painting of a bird on a table with a bird on it.
ZeroCap: A night with Vincent.
Ours:
A famous Gogh painting after streaming moonlight over all the grand structures.
A view despite a nocturnal sky within famous mainstream artworks.
A nighttime sky can appear in drawings and oil paintings.

GRIT: A busy city street with lots of people walking on.
CLIPCap:A busy city street with people crossing it.
ViTCap: a city street with people walking and a bus.
ZeroCap:A billboard in the middle of of of a busy intersection.
Ours:
A new york time square.
A busy billboard covered time square with scenery.
A landscape masking time square depicting a vibrant morning.
A city featuring yellow billboard and advertisements on google outdoor.

Figure 6. Results of various image styles.

VERB ADV . Results are shown in Table. 4.

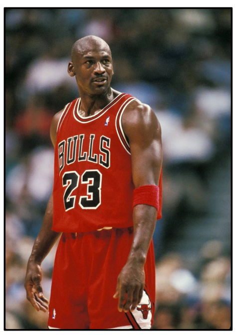
F. Bad case analysis

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.

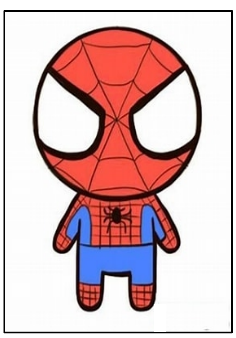
GRIT: a very tall tower with a clock tower in the
CLIPCap: A tall tower with a clock on top.
ViTCap:a picture of a tall tower with a clock on it.
ZeroCap:Image of a French Italian landmark is captioned on the first of the month.

Ours:
image of a famous landmark in france.
image of a tall structure iconic in famous french photographs.



GRIT:a man in a red uniform holding a basketball
CLIPCap:A man wearing a red neck tie holding a ball.
ViTCap:a close up of a baseball player holding a ball
ZeroCap:Image of a hero NBA All-Star Michael Jordan in Sports Illustrated uniform

Ours:
A Jordan steel bulls jersey taken from NBA promotional material.
A 1990s superstar wearing bright rose gold jerseys.



GRIT: a red teddy bear with a heart on a
CLIPcap:A Star Wars character is wearing a star trek neck tie.
ViTcap:a drawing of a man holding a giant shark.
Zerocap: Image of a character from Spider-Man comic art.

Ours:
A minime for super spiderman animated.
A stylized spider wearing a small comic outfit.

Figure 7. Results of images containing world knowledge.