

# Supplementary Materials: See or Guess: Counterfactually Regularized Image Captioning

Anonymous Authors

**Table 1: Evaluation results of different decoding algorithms on CHAIR<sub>s</sub> values on Flickr30k Entities and MSCOCO.**

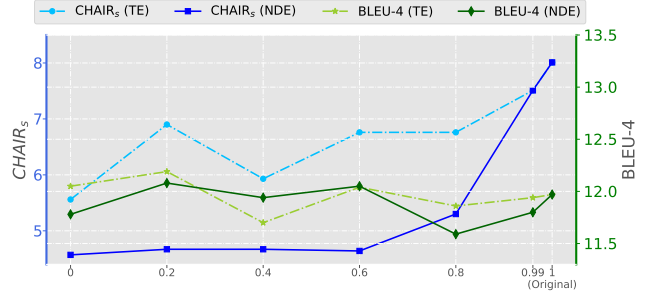
Methods	Flickr30k Entities				MSCOCO			
	Beam	Greedy	TopK	Nucleus	Beam	Greedy	TopK	Nucleus
<b>ClipCap</b>	20.45	19.82	15.06	15.55	64.05	66.43	56.00	62.45
+ObjL	21.18	20.34	16.28	16.49	64.64	65.48	56.80	61.81
+ObjMLM	25.37	24.54	19.12	19.89	70.07	74.20	65.56	67.99
+TE (ours)	19.78	19.29	14.92	<b>14.57</b>	63.58	64.97	54.73	59.09
+NDE (ours)	<b>19.64</b>	<b>19.05</b>	<b>14.60</b>	<u>14.99</u>	<b>63.04</b>	<b>64.22</b>	<b>52.71</b>	<b>58.27</b>
<b>BLIP</b>	12.14	12.01	9.51	8.51	33.70	35.04	30.17	31.04
+ObjL	10.61	11.52	9.41	7.51	33.07	32.56	27.86	30.08
+ObjMLM	<u>10.11</u>	11.01	7.71	7.29	33.90	36.45	30.70	31.26
+TE (ours)	10.23	<u>10.92</u>	<u>7.60</u>	<u>6.81</u>	<u>31.10</u>	<u>32.02</u>	<u>27.71</u>	<u>28.40</u>
+NDE (ours)	<b>9.53</b>	<b>10.51</b>	<b>7.51</b>	<b>6.80</b>	<b>30.43</b>	<b>31.04</b>	<b>26.97</b>	<b>27.86</b>
<b>BLIP2</b>	8.01	<u>7.60</u>	7.81	6.52	30.28	<u>28.46</u>	22.58	26.21
+ObjL	8.02	7.74	7.08	7.46	30.26	29.45	22.20	25.51
+ObjMLM	8.12	7.64	7.57	7.67	34.84	32.00	25.57	29.52
+TE (ours)	<u>7.61</u>	<b>7.39</b>	<u>6.24</u>	<b>6.10</b>	<u>29.60</u>	<b>28.08</b>	<u>22.14</u>	<b>23.64</b>
+NDE (ours)	<b>7.51</b>	<b>7.39</b>	<b>6.21</b>	<u>6.17</u>	<b>29.26</b>	<b>28.08</b>	<b>21.99</b>	<u>24.22</u>

## A EVALUATION ON DECODING ALGORITHMS

Our method penalizes the occurrence probability of specific tokens, which may raise concerns regarding reliance on the generation algorithm. To address this, we conducted experiments employing different decoding algorithms, e.g., greedy search, top-K sampling, and nucleus sampling, besides the beam search strategy mentioned before. In our experiments, we set beam=5 for beam search, K=10 for top-K sampling, and p=0.8 for nucleus sampling. As presented in Table 1, the results on CHAIR<sub>s</sub> demonstrate that our methods consistently outperform the baselines, regardless of the decoding algorithm. It is worth noting that our primary focus lies in examining the distinctions among various methods within the same decoding algorithm, rather than emphasizing the differences between different decoding algorithms. Thus we use adopt beam search as our general setting in our previous sections. This observation highlights the robustness of our approaches across various decoding algorithms, further proving their effectiveness.

## B IMPACT OF $\alpha$

We further investigate the impact of the hyperparameter  $\alpha$  on producing object hallucination. BLIP2 is adopted as the backbone and the results are depicted in Figure 1. Generally speaking, CHAIR<sub>s</sub> gradually rises as  $\alpha$  increases. This makes sense since either our TE or NDE methods serve as a regularization. The less the regularization, the worse the result. It experiences substantial alterations while the parameter alpha ranges between 0.8 and 1. However, when  $\alpha$  gradually converges to 1, the model degenerates into a vanilla training process. In addition, we find it no significant drops in model generation performance on factual images as  $\alpha$  decreases.



**Figure 1: The variation line chart of CHAIR<sub>s</sub> and BLEU-4 of BLIP2 when  $\alpha$  changes on TE and NDE. The value of CHAIR<sub>s</sub> shows a clear change trend, while the value of BLEU-4 fluctuates insignificantly. Best viewed in color.**

This evidence substantiates the assertion that our approach maintains the model’s performance intact in factual scenarios.

## C IMPLEMENTATION DETAILS

As previously elucidated, our optimization contains two stages. In stage one, ClipCap and BLIP are trained on Flickr30k Entities and MSCOCO for 10 epochs with a learning rate at 5e-5/1e-5, and the batch size is set to 128/32, respectively. As for BLIP2, the learning rate is respectively set to 1e-6/7e-6 for Flickr and MSCOCO, and it is trained for 5 epochs with a batch size 64. In stage two, the TE/NDE loss is added for 2 epochs until the aggregate loss on validation converges, with the hyperparameter  $\alpha$  set to 1-1e-3/1-1e-8 on Flickr30k Entities and 1-1e-4/1-3e-5 on MSCOCO for ClipCap, while for BLIP2 it is 1-9e-3/1-9e-3 on Flickr30k Entities and 1-5e-4/1-6e-4 on MSCOCO, and 1-1e-4/1-1e-2 on Flickr30k Entities and 1-1e-4 on MSCOCO for BLIP2, respectively. For BLIP2, we adopt ViT-L and OPT-2.7b as the visual encoder and the language model, which are frozen during training. We use beam search for all backbones with a beam size of 5 and a maximum length of 20 during inference.

## D MORE CASES

We present more cases of masked counterfactual images and inpainted counterfactual images across different methods, displayed in Figure 2 and 3. Through the comparison of various methodologies, our approach consistently generates image captions that exhibit greater fidelity to the underlying visual content. By avoiding unreliable conjectures, our methods can successfully mitigate the occurrence of object hallucinations, thereby augmenting the robustness and reliability of various image captioning models. Nevertheless, all methods may still encounter challenges in some complex situations. A comprehensive analysis and subsequent improvements are necessary to enhance both reliability and validity in future investigations.



GT: A dog carries an object through the snowy grass.  
 ClipCap: A brown dog is carrying **a large stick** in its mouth.  
 ClipCap+ObjL: A dog is running through the snow **carrying a stick** in its mouth.  
 ClipCap+ObjMLM: A brown dog is running through the snow **with a stick in his mouth**.  
 ClipCap+TE: A brown dog is climbing a snowy hill.  
 ClipCap+NDE: A dog leaps through the snow.



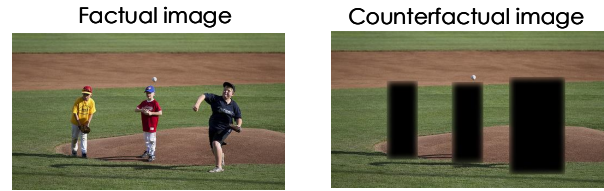
GT: People are riding thrill game.  
 BLIP: **A man in a red shirt is climbing** a red structure.  
 BLIP+ObjL: **A man in a white shirt is standing on** a red structure with trees in the background.  
 BLIP+ObjMLM: **A man in a white shirt is climbing** a red structure.  
 BLIP+TE: A red playground set with a tree in the background.  
 BLIP+NDE: A red play set in a park with trees in the background.



GT: A distant airplane flying between two large buildings.  
 BLIP2: Two tall buildings **with a boat** in front of them.  
 BLIP2+ObjL: Two tall buildings **with a kite flying** in the background.  
 BLIP2+ObjMLM: Two tall buildings **with a plane** in the sky.  
 BLIP2+TE: A city with two tall buildings in the background.  
 BLIP2+NDE: A city with two tall buildings in the background.



GT: An explorer is jumping for joy near his snow bicycle at the edge of a large body of water in a area covered with snow.  
 ClipCap: **A man is doing a trick on his snowboard**.  
 ClipCap+ObjL: **A man is doing a trick on a snowboard**.  
 ClipCap+ObjMLM: **A man is doing a trick** on a snow covered hillside with the wind blowing in the background.  
 ClipCap+TE: **A man is jumping** over a snow covered hill.  
 ClipCap+NDE: **A man is jumping** over a snow covered hill.



GT: Three children playing baseball in uniforms on a baseball diamond.  
 BLIP: **A man is throwing** a baseball on a baseball field.  
 BLIP+ObjL: **A baseball player about to throw a ball** on a baseball field.  
 BLIP+ObjMLM: **A baseball player about to throw** the ball.  
 BLIP+TE: A baseball is in the middle of a pitch.  
 BLIP+NDE: A baseball in the middle of a field.



GT: A man holding a giant pair of black scissors.  
 BLIP2: A man in a room with **a blank wall behind him**.  
 BLIP2+ObjL: A man holding a large object.  
 BLIP2+ObjMLM: A man holding a large black object.  
 BLIP2+TE: A man with glasses and a plaid shirt.  
 BLIP2+NDE: A man with glasses and a plaid shirt.

Figure 2: Some examples of generated captions by various methods for some masked counterfactual images. Phrases highlighted in red are hallucinations that do not exist in the counterfactual image.

Factual image



Counterfactual image



GT: A small dog is standing behind a camera.

In masked scenario:

BLIP: A camera with a flash attached to it.

BLIP+ObjL: A camera with a lens attached to it's body.

BLIP+ObjMLM: A camera with a lens attached to it.

BLIP+TE: A camera with a flash attached to it.

BLIP+NDE: A camera with a flash attached to it.

In inpainted scenario:

BLIP: A person is cleaning the floor with a broom.

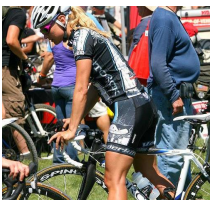
BLIP+ObjL: A camera with a broom on top of it.

BLIP+ObjMLM: A person is cleaning the floor with a broom.

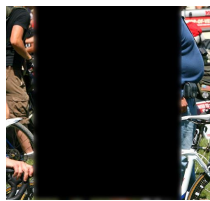
BLIP+TE: A camera with a broom on top of it.

BLIP+NDE: A camera with a broom on top of it.

Factual image



Counterfactual image



GT: Blond woman in black cycling outfit and bicycle helmet getting on a ten speed bike.

In masked scenario:

ClipCap: A woman in a pink shirt is riding a bicycle.

ClipCap+ObjL: A woman in a white tank top is riding a bike.

ClipCap+ObjMLM: A woman in a white tank top and black shorts is riding a bike.

ClipCap+TE: A group of bicyclists in a field of flowers.

ClipCap+NDE: A woman in a black tank top is riding a bike.

In inpainted scenario:

ClipCap: A white dog with a red collar is riding on a bicycle.

ClipCap+ObjL: A group of people are watching a dog on a red leash.

ClipCap+ObjMLM: A dog with a red collar is on a red and white bicycle with a man in the background.

ClipCap+TE: A dog is standing on a bicycle in front of a crowd of people.

ClipCap+NDE: A black and white dog rides on a bicycle.

Factual image



Counterfactual image



GT: A young boy swings a baseball bat at a ball in the park.

In masked scenario:

BLIP2: A man throwing a frisbee in a field.

BLIP2+ObjL: A person playing frisbee in a field.

BLIP2+ObjMLM: Two people playing frisbee in a field.

BLIP2+TE: A man throwing a frisbee in a field.

BLIP2+NDE: Two people playing frisbee in a field.

In inpainted scenario:

BLIP2: Two children playing frisbee in a field.

BLIP2+ObjL: A child playing with a frisbee.

BLIP2+ObjMLM: A little boy playing with a frisbee.

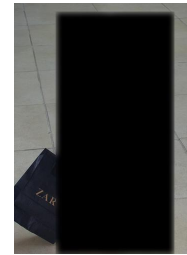
BLIP2+TE: Two people playing frisbee in a field.

BLIP2+NDE: Two people playing frisbee in a field.

Factual image



Counterfactual image



GT: A young girl is carrying two large black shopping bags.

In masked scenario:

BLIP2: A person sitting on the floor with a black bag on the floor next to them.

BLIP2+ObjL: A black bag with zara written on it.

BLIP2+ObjMLM: A black bag on the floor next to a person's foot.

BLIP2+TE: A person sitting on the floor with a black bag on the floor next to them.

BLIP2+NDE: A person sitting on the floor with a black bag on the floor next to them.

In inpainted scenario:

BLIP2: A metal gear on the floor next to a credit card.

BLIP2+ObjL: A metal gear on the floor next to a card with zara written on it.

BLIP2+ObjMLM: A piece of metal with gears on it.

BLIP2+TE: A metal gear on the floor next to a credit card.

BLIP2+NDE: A metal gear on the floor next to a bag with zara written on it.

Figure 3: Some examples of generated captions by various methods for some masked and inpainted counterfactual images. Phrases highlighted in red are hallucinations that do not exist in the counterfactual image.