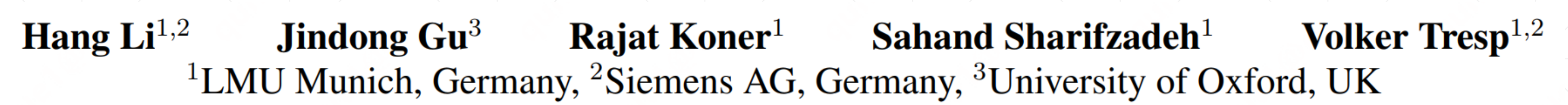
arxiv

Do DALL-E and Flamingo Understand Each Other?

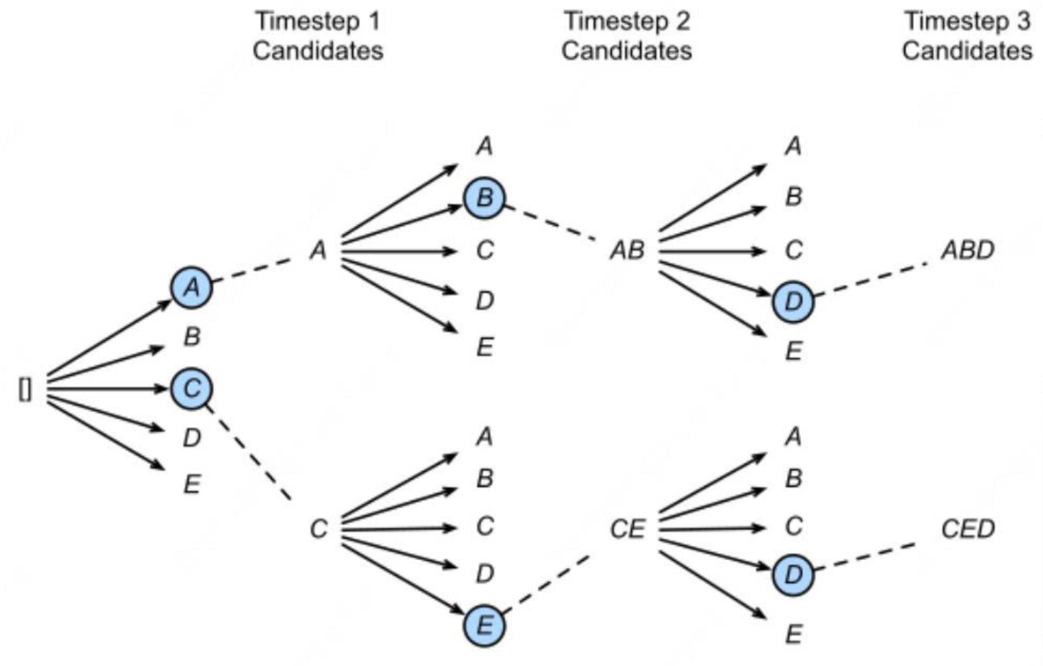


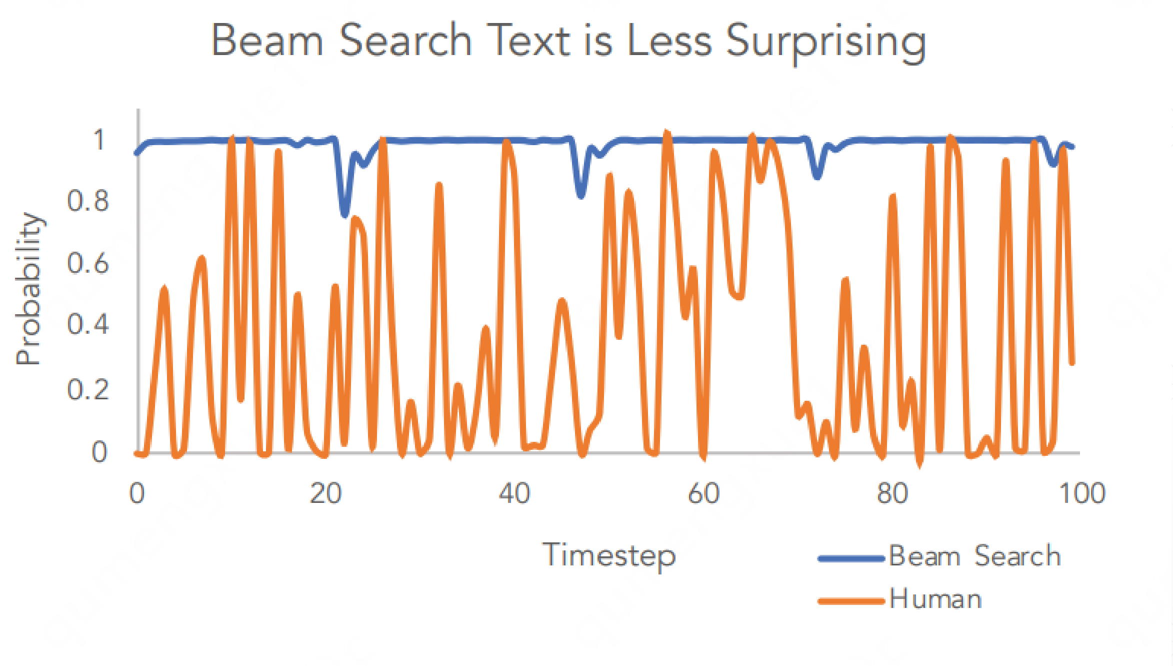
Hang Li[1,2] **Jindong Gu**[3] **Rajat Koner**[1] **Sahand Sharifzadeh**[1] **Volker Tresp**[1,2]

[1]LMU Munich, Germany, [2]Siemens AG, Germany, [3]University of Oxford, UK

Mengxue

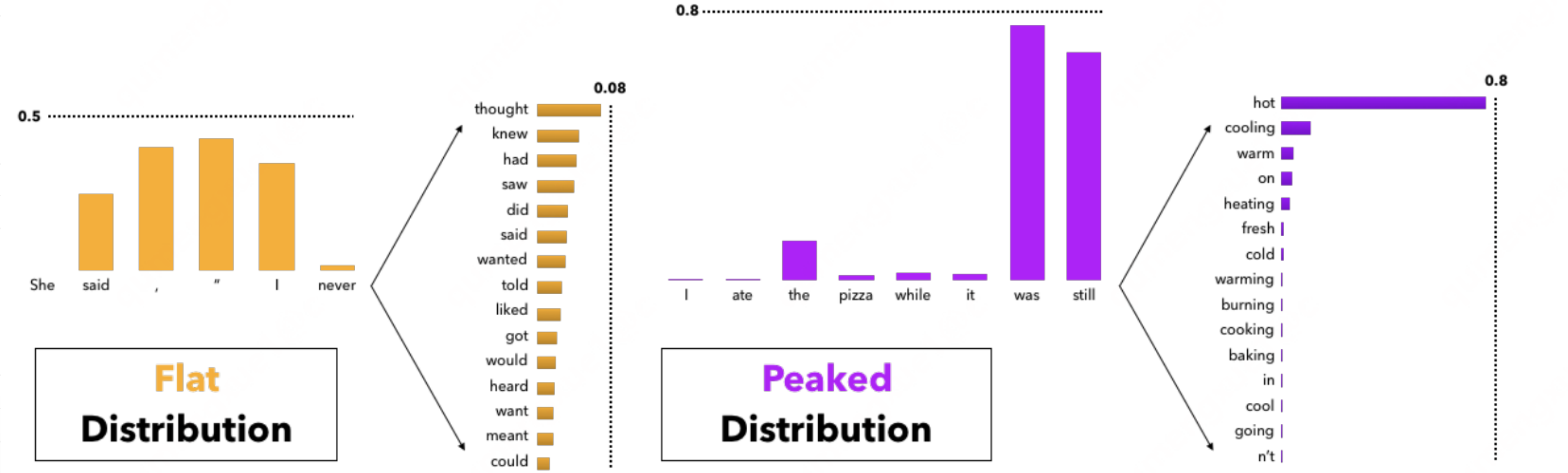
Preview: Beam Search





Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." *ICLR 2020*.

Preview: Top-k Sampling

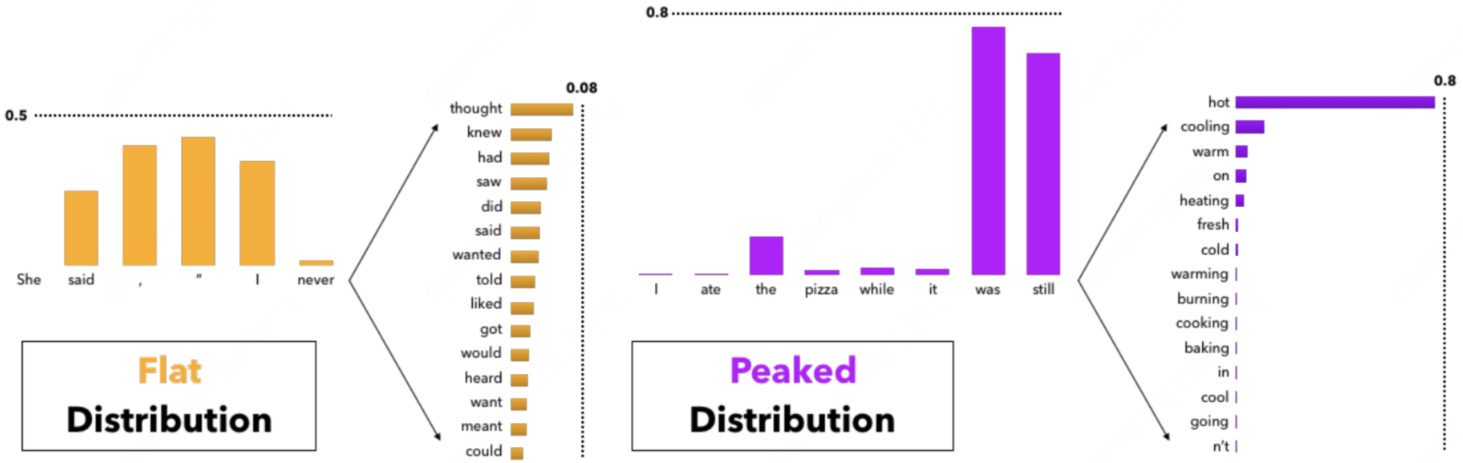


- 这个方法就是在采样前将输出的概率分布截断，取出概率最大的k个词构成一个集合，然后将这个子集词的概率再归一化，最后从新的概率分布中采样词汇。

- 但因为概率分布变化比较大，有时候可能很均匀(flat)，有的时候比较集中(peaked)。当分布均匀时，一个较小的k容易丢掉很多优质候选词。但如果k定的太大，这个方法会退化回全局随机采样。

Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." *ICLR 2020.*

Preview: Nucleus sampling(Top-p Sampling)



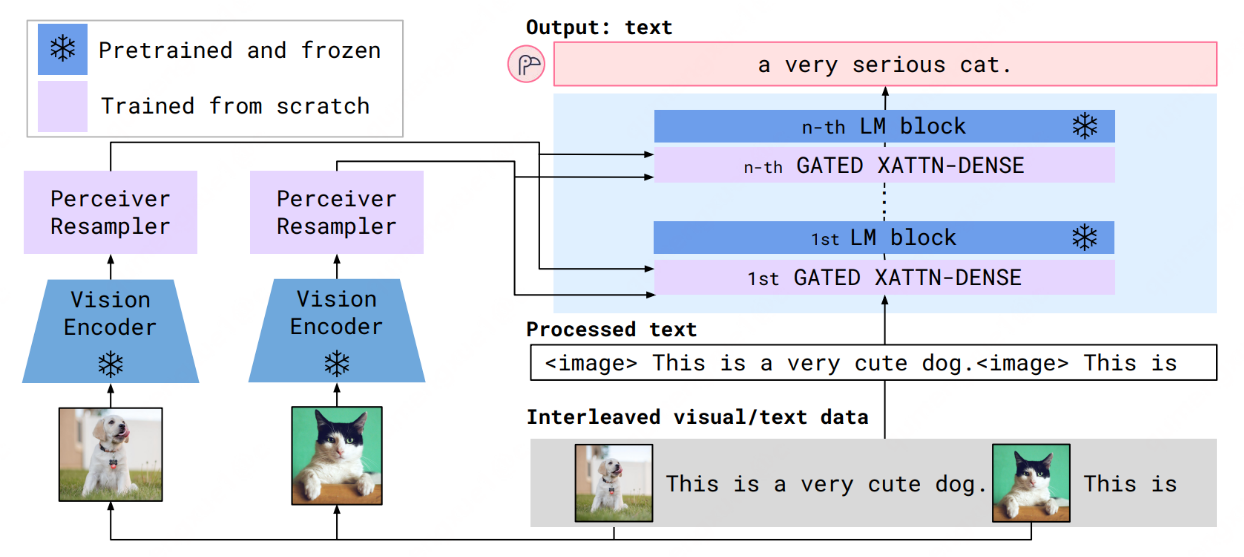
- 在每个时间步，解码词的概率分布满足80/20原则或者说长尾分布，头部的几个词的出现概率已经占据了绝大部分概率空间，把这部分核心词叫做nucleus。
- 基于这样的观察，提出nucleus sampling：给定一个概率阈值p，从解码词候选集中选择一个最小集Vp，使得它们出现的概率和大于等于p。然后再对Vp做一次re-scaling，本时间步仅从Vp集合中解码。
- 这样的好处在于在不同时间步，随着解码词的概率分布不同，候选词集合的大小会动态变化，不像top-k sampling是一个固定的窗口大小。由于解码词还是从头部候选集中筛选，这样的动态调整可以使生成的句子在满足多样性的同时又保持通顺。

Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." *ICLR 2020.*

What’ s the best choice?

- Beam Search → Top-k Sampling → Top-p Sampling
- **Caption** → Given an image, what textual description most accurately describes the image?
- **Generation** → Given a text, what is the best image that can present the semantics of the text?

Do DALL-E and Flamingo Understand Each Other?



(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

Image Captioning

Text-to-Image Generation

Flammingo

DALLE-E

Motivation

- In this work, we argue that the best text or caption for a given image is the text which would generate the image which is the most similar to that image.

Cycle Consistency

- Likewise, the best image for a given text is the image that results in the caption which is best aligned with the original text.

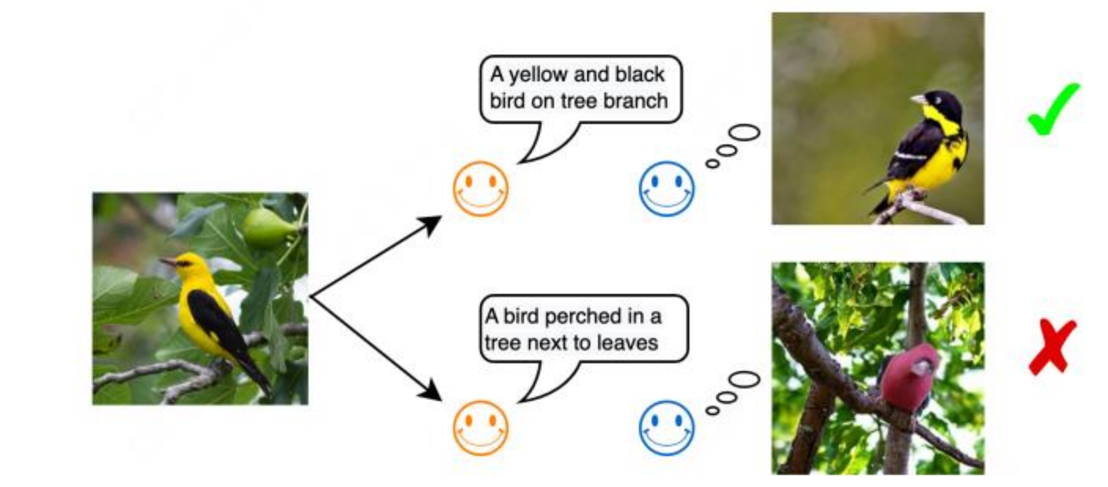
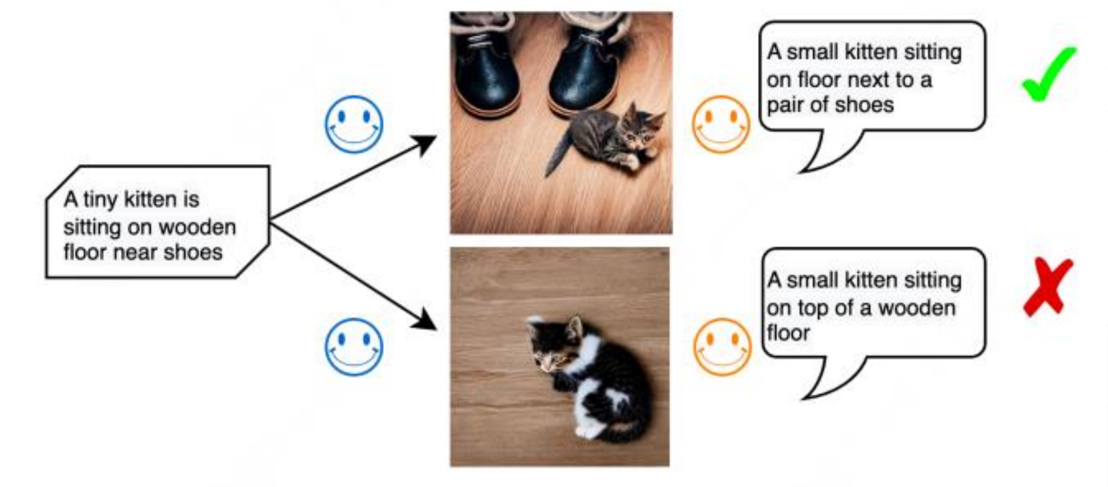

A yellow and black bird on tree branch
A bird perched in a tree next to leaves

Image Captioning


A tiny kitten is sitting on wooden floor near shoes
A small kitten sitting on floor next to a pair of shoes
A small kitten sitting on top of a wooden floor

Text-to-Image Generation

Framework

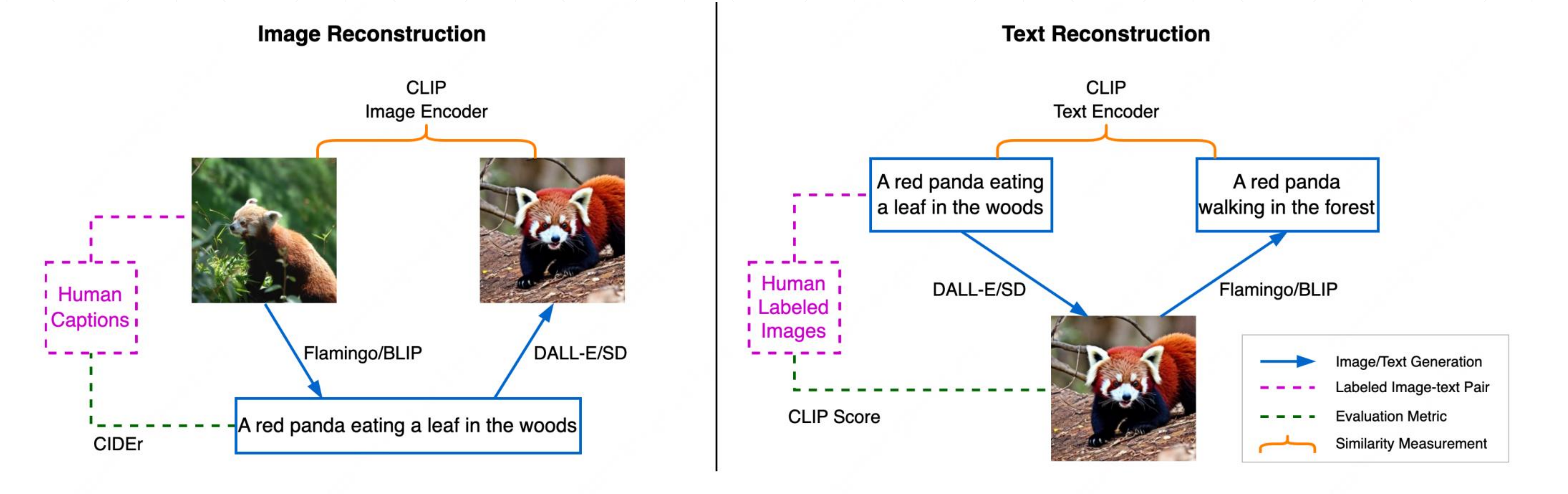


Image Reconstruction
CLIP
Image Encoder
Human
Captions
Flamingo/BLIP
DALL-E/SD
A red panda eating a leaf in the woods
CIDEr
Text Reconstruction
CLIP
Text Encoder
A red panda eating
a leaf in the woods
A red panda
walking in the forest
Human
Labeled
Images
DALL-E/SD
Flamingo/BLIP
CLIP Score
Image/Text Generation
Labeled Image-text Pair
Evaluation Metric
Similarity Measurement

Experiments

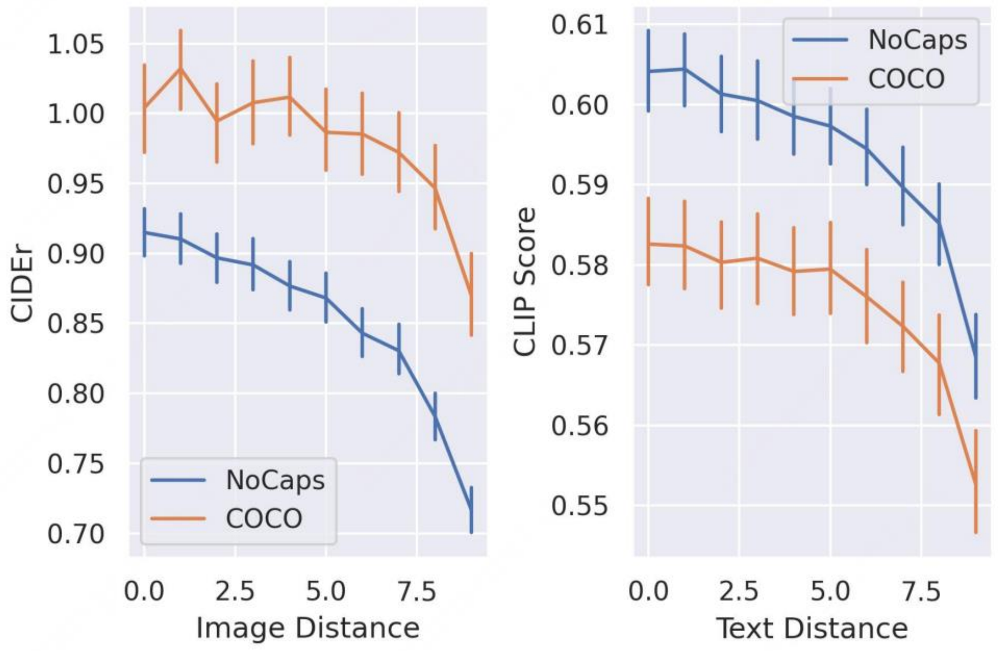
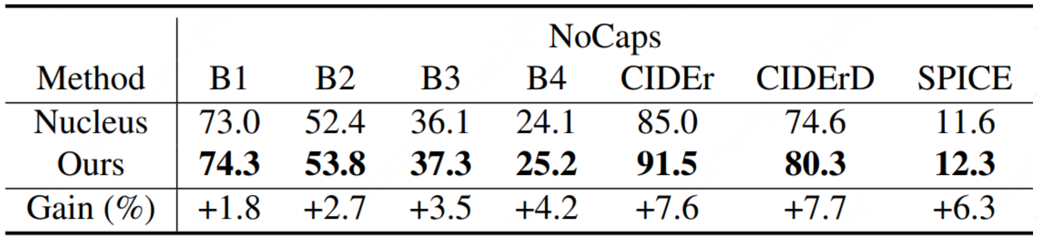
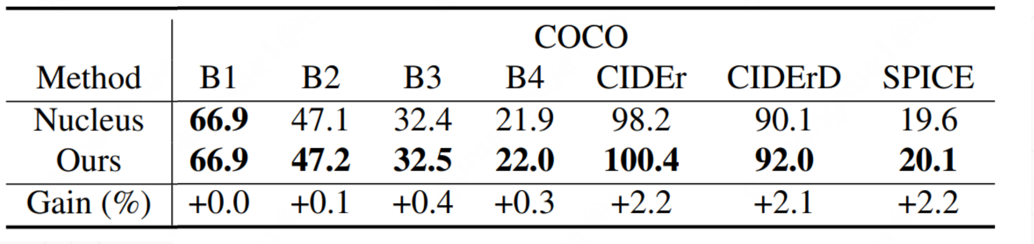


Figure 3. Left: for each given image, the better the reconstructed image (shown in x-axis), the better the caption (shown in y-axis). Right: for each given text, the better the reconstructed text (shown in x-axis), the better the image (shown in y-axis).

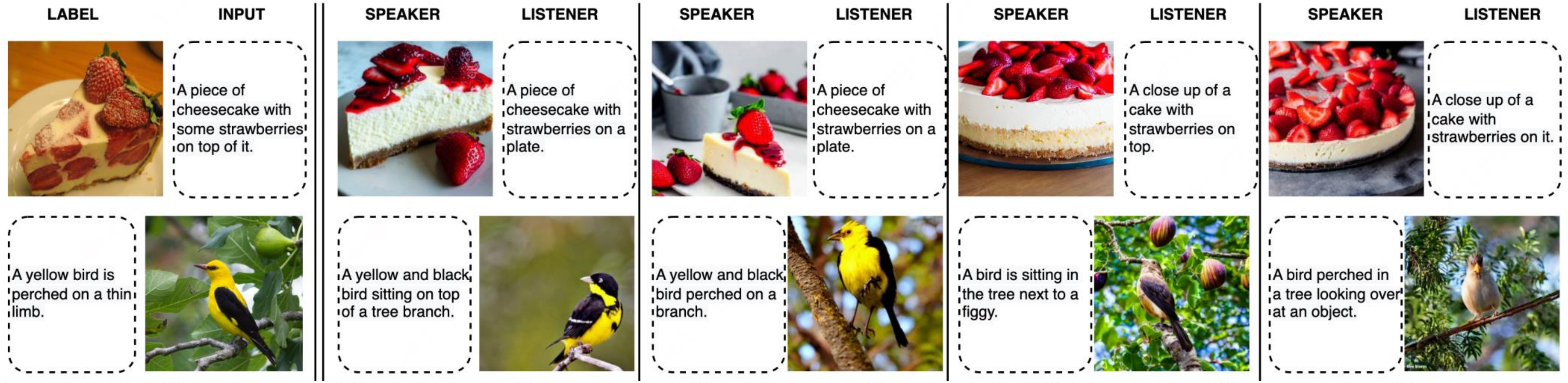


	NoCaps						
Method	B1	B2	B3	B4	CIDEr	CIDErD	SPICE
Nucleus	73.0	52.4	36.1	24.1	85.0	74.6	11.6
Ours	**74.3**	**53.8**	**37.3**	**25.2**	**91.5**	**80.3**	**12.3**
Gain (%)	+1.8	+2.7	+3.5	+4.2	+7.6	+7.7	+6.3



	COCO						
Method	B1	B2	B3	B4	CIDEr	CIDErD	SPICE
Nucleus	**66.9**	47.1	32.4	21.9	98.2	90.1	19.6
Ours	**66.9**	**47.2**	**32.5**	**22.0**	**100.4**	**92.0**	**20.1**
Gain (%)	+0.0	+0.1	+0.4	+0.3	+2.2	+2.1	+2.2

Qualitative Results



A piece of cheesecake with some strawberries on top of it.

A piece of cheesecake with strawberries on a plate.

A piece of cheesecake with strawberries on a plate.

A close up of a cake with strawberries on top.

A close up of a cake with strawberries on it.

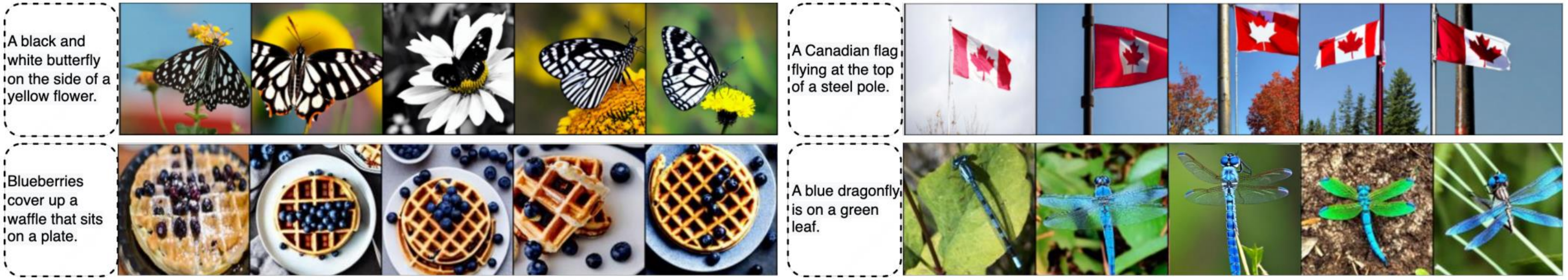
A yellow bird is perched on a thin limb.

A yellow and black bird sitting on top of a tree branch.

A yellow and black bird perched on a branch.

A bird is sitting in the tree next to a figgy.

A bird perched in a tree looking over at an object.



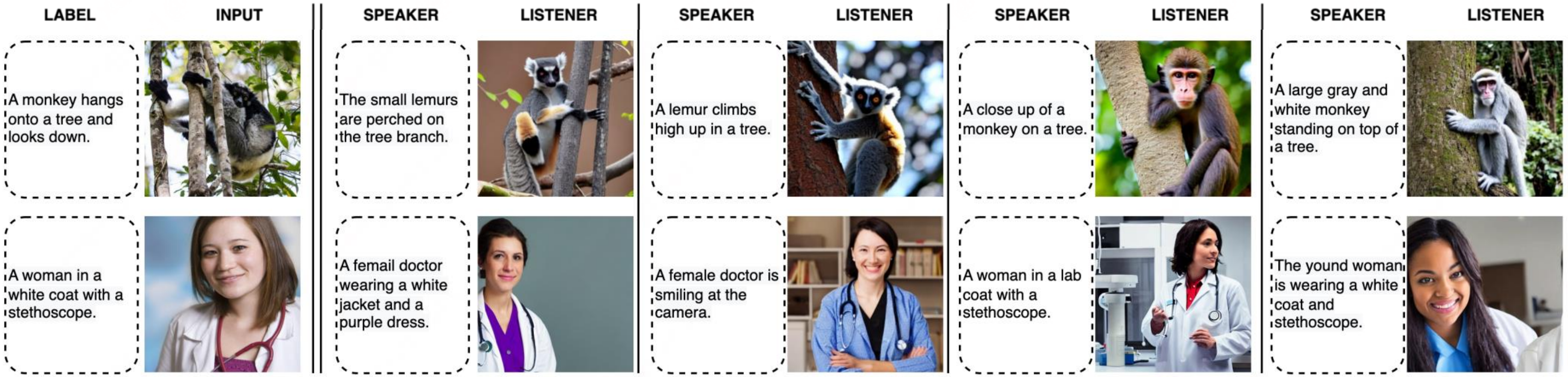
A black and white butterfly on the side of a yellow flower.

A Canadian flag flying at the top of a steel pole.

Blueberries cover up a waffle that sits on a plate.

A blue dragonfly is on a green leaf.

Failure Cases



A monkey hangs onto a tree and looks down.

The small lemurs are perched on the tree branch.

A lemur climbs high up in a tree.

A close up of a monkey on a tree.

A large gray and white monkey standing on top of a tree.

A woman in a white coat with a stethoscope.

A femail doctor wearing a white jacket and a purple dress.

A female doctor is smiling at the camera.

A woman in a lab coat with a stethoscope.

The yound woman is wearing a white coat and stethoscope.

Annotation Imperfection

Generation Bias

Thanks