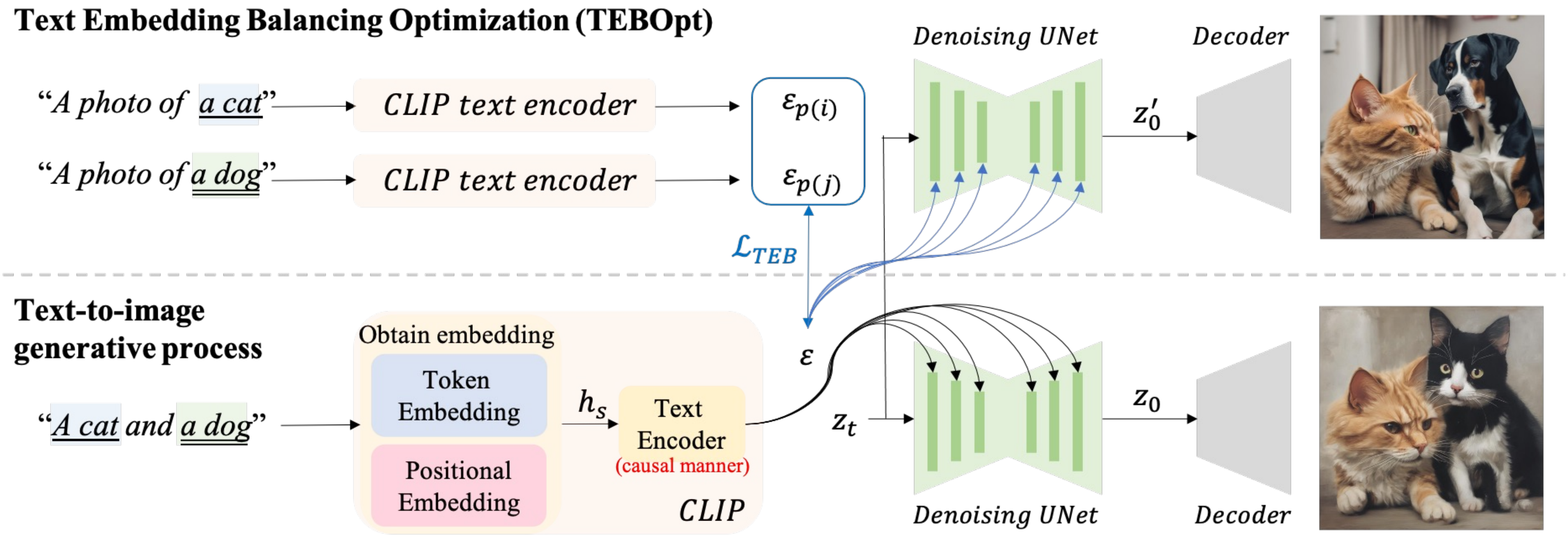
NEURAL INFORMATION PROCESSING SYSTEMS 2024

國立陽明交通大學 NATIONAL YANG MING CHIAO TUNG UNIVERSITY

Georgia Tech

A Cat Is A Cat (Not A Dog!): Unraveling Information Mix-ups in Text-to-Image Encoders through Causal Analysis and Embedding Optimization

Chieh-Yun Chen, Chiang Tseng, Li-Wu Tsao and Hong-Han Shuai



Motivation

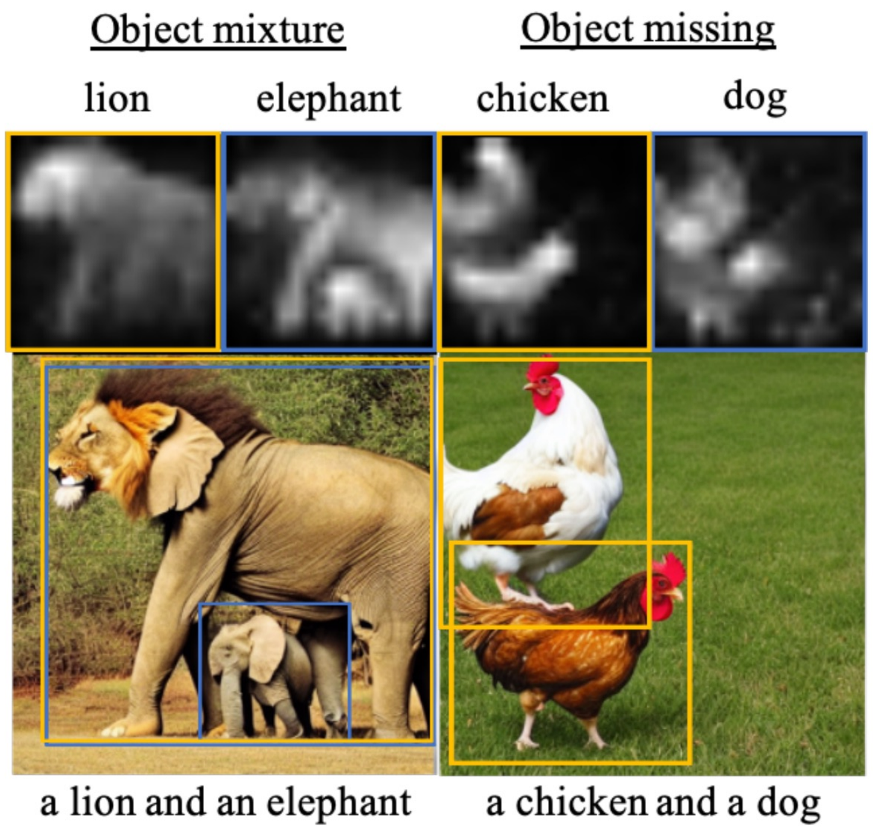


Figure 1: Visualization of cross-attention maps when object mixture and missing occur.

Information bias towards the first mentioned object

Prompt	(a) A/An <obj1> and a/an <obj2>	(b) A/An <obj2> and a/an <obj1>
2 objects exist	12.25%	11.75%
mixtures	20.25%	18.75%
only obj1 exist	46.00%	21.75%
only obj2 exist	20.00%	47.00%
no target object	1.50%	0.75%
Info bias	2.30	0.46

Table 1: Both prompts strongly bias towards the first mentioned object. The bias generally exists in more objects, reported in Supplement D.

Motivation

Figure 1: Visualization of cross-attention maps when object mixture and missing occur.

Information bias towards the first mentioned object

Prompt	(a) A/An <obj1> and a/an <obj2>	(b) A/An <obj2> and a/an <obj1>
2 objects exist	12.25%	11.75%
mixtures	20.25%	18.75%
only obj1 exist	46.00%	21.75%
only obj2 exist	20.00%	47.00%
no target object	1.50%	0.75%
Info bias	2.30	0.46

Table 1: Both prompts strongly bias towards the first mentioned object. The bias generally exists in more objects, reported in Supplement D.

Motivation

Figure 1: Visualization of cross-attention maps when object mixture and missing occur.

Information bias towards the first mentioned object

Prompt	(a) A/An <obj1> and a/an <obj2>	(b) A/An <obj2> and a/an <obj1>
2 objects exist	12.25%	11.75%
mixtures	20.25%	18.75%
only obj1 exist	46.00%	21.75%
only obj2 exist	20.00%	47.00%
no target object	1.50%	0.75%
Info bias	2.30	0.46

Table 1: Both prompts strongly bias towards the first mentioned object. The bias generally exists in more objects, reported in Supplement D.

Motivation

a lion and an elephant   a chicken and a dog

Figure 1: Visualization of cross-attention maps when object mixture and missing occur.

Information bias towards the first mentioned object

Prompt	(a) A/An <obj1> and a/an <obj2>	(b) A/An <obj2> and a/an <obj1>
2 objects exist	12.25%	11.75%
mixtures	20.25%	18.75%
only obj1 exist	46.00%	21.75%
only obj2 exist	20.00%	47.00%
no target object	1.50%	0.75%
Info bias	2.30	0.46

Table 1: Both prompts strongly bias towards the first mentioned object. The bias generally exists in more objects, reported in Supplement D.

Motivation

Figure 1: Visualization of cross-attention maps when object mixture and missing occur.

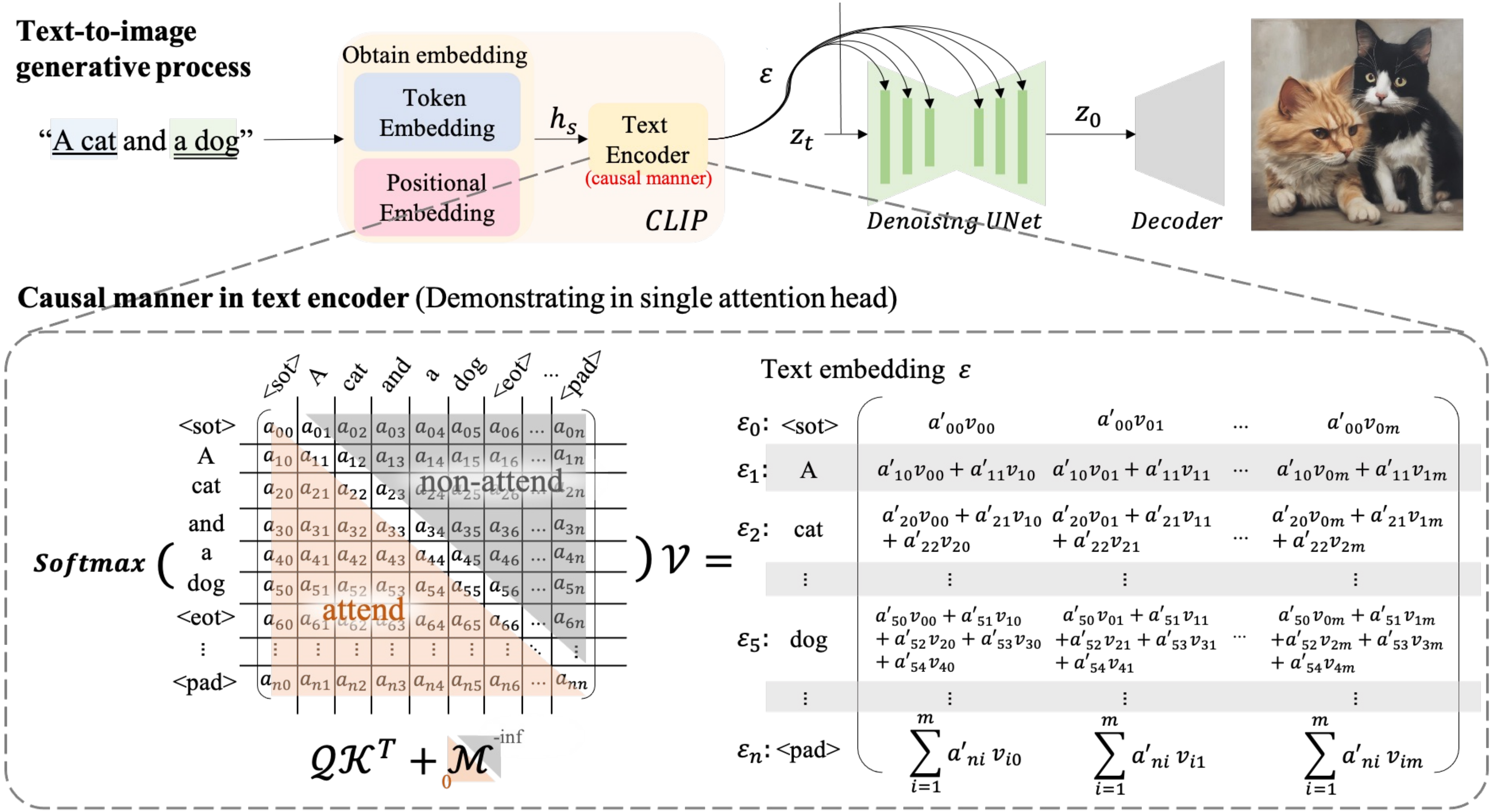
Information bias towards the first mentioned object

Prompt	(a) A/An <obj1> and a/an <obj2>	(b) A/An <obj2> and a/an <obj1>
2 objects exist	12.25%	11.75%
mixtures	20.25%	18.75%
only obj1 exist	46.00%	21.75%
only obj2 exist	20.00%	47.00%
no target object	1.50%	0.75%
Info bias	2.30	0.46

Table 1: Both prompts strongly bias towards the first mentioned object. The bias generally exists in more objects, reported in Supplement D.

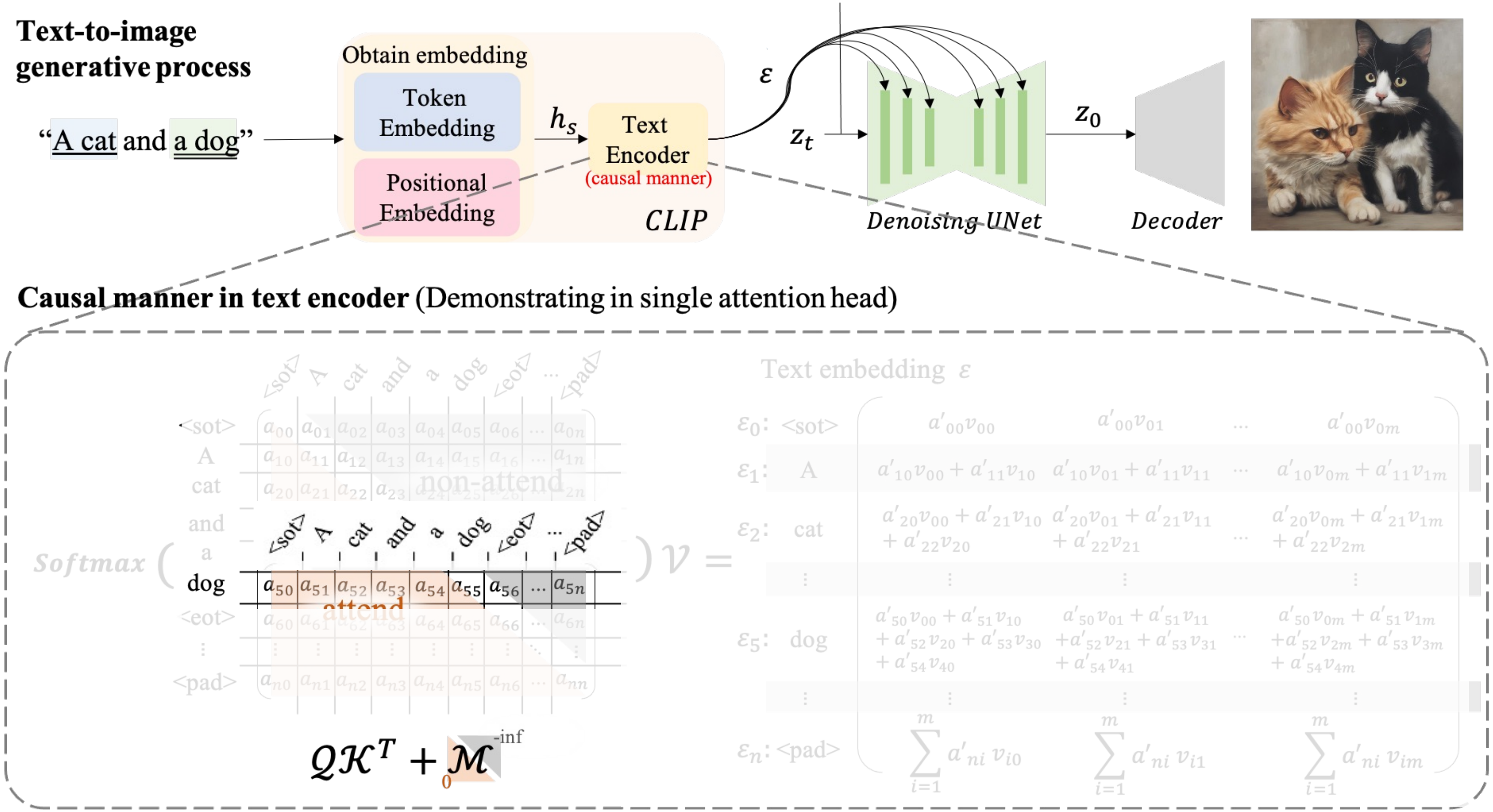
Causal manner leads to information bias

Information bias towards the first mentioned object

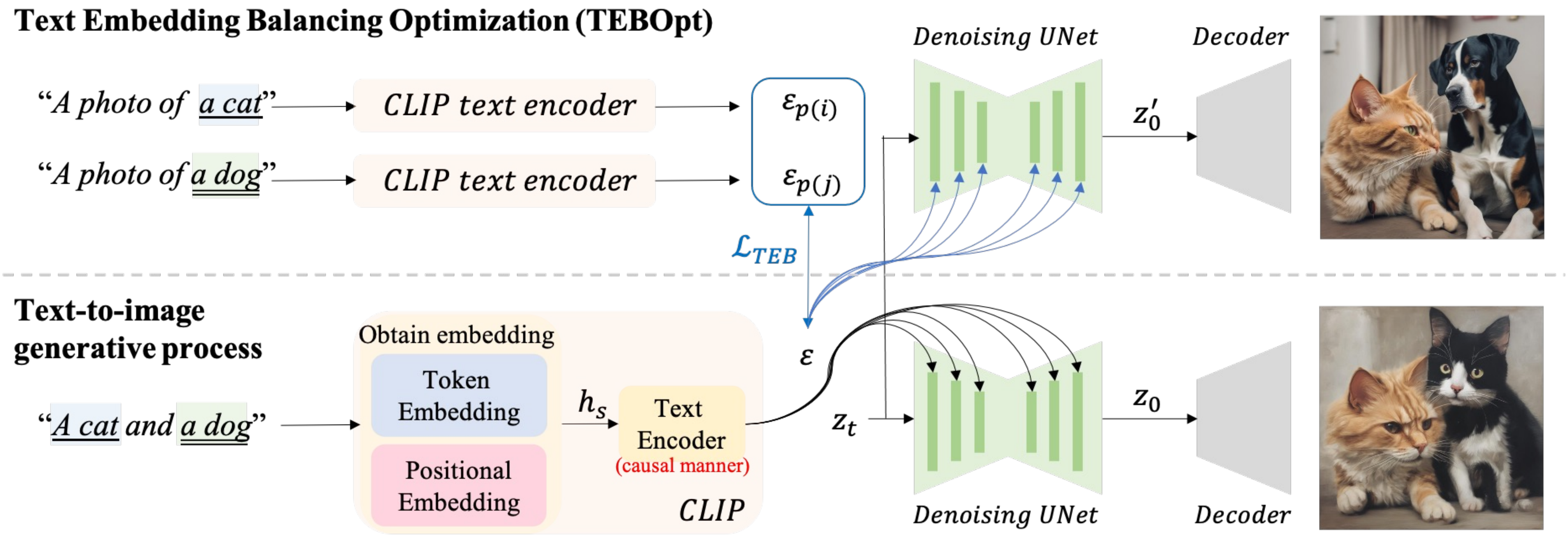


Causal manner leads to information bias

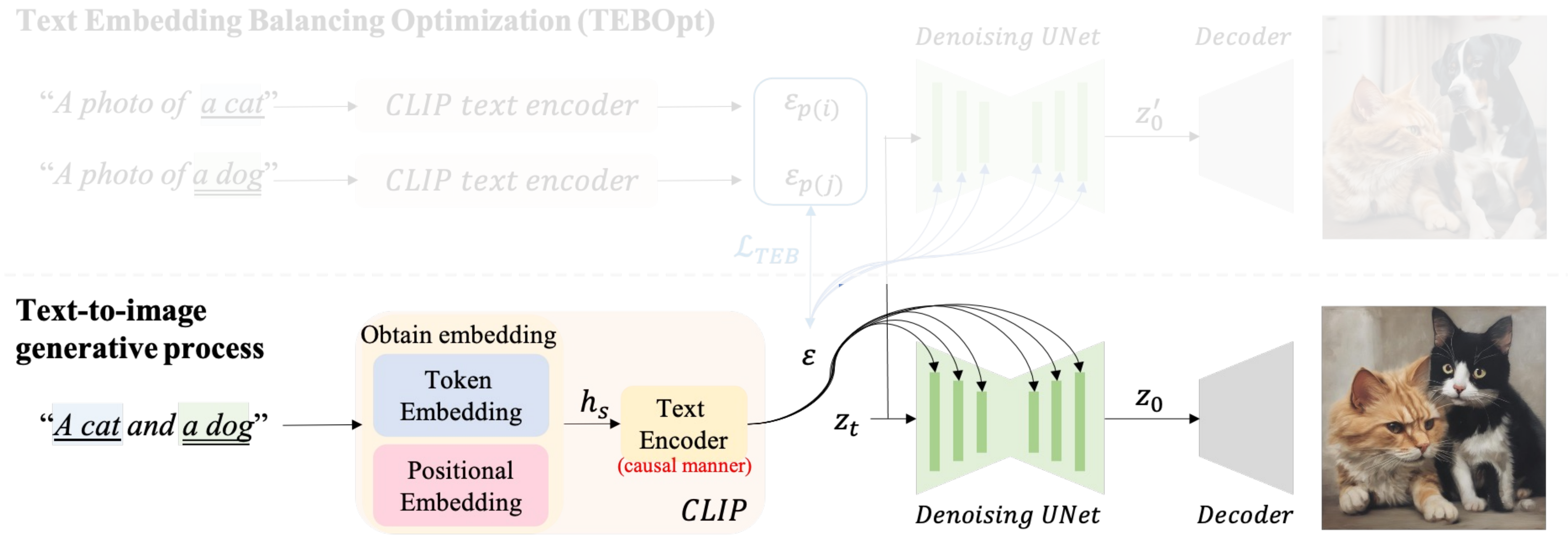
Information bias towards the first mentioned object



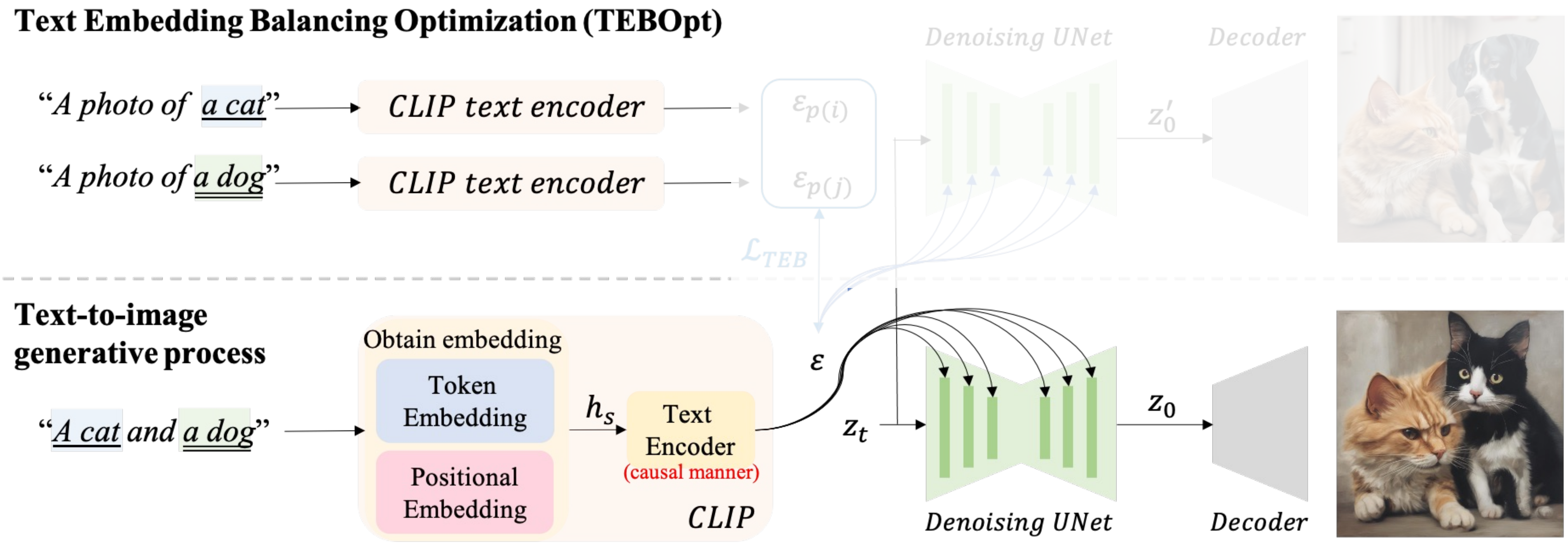
The proposed text embedding optimization method



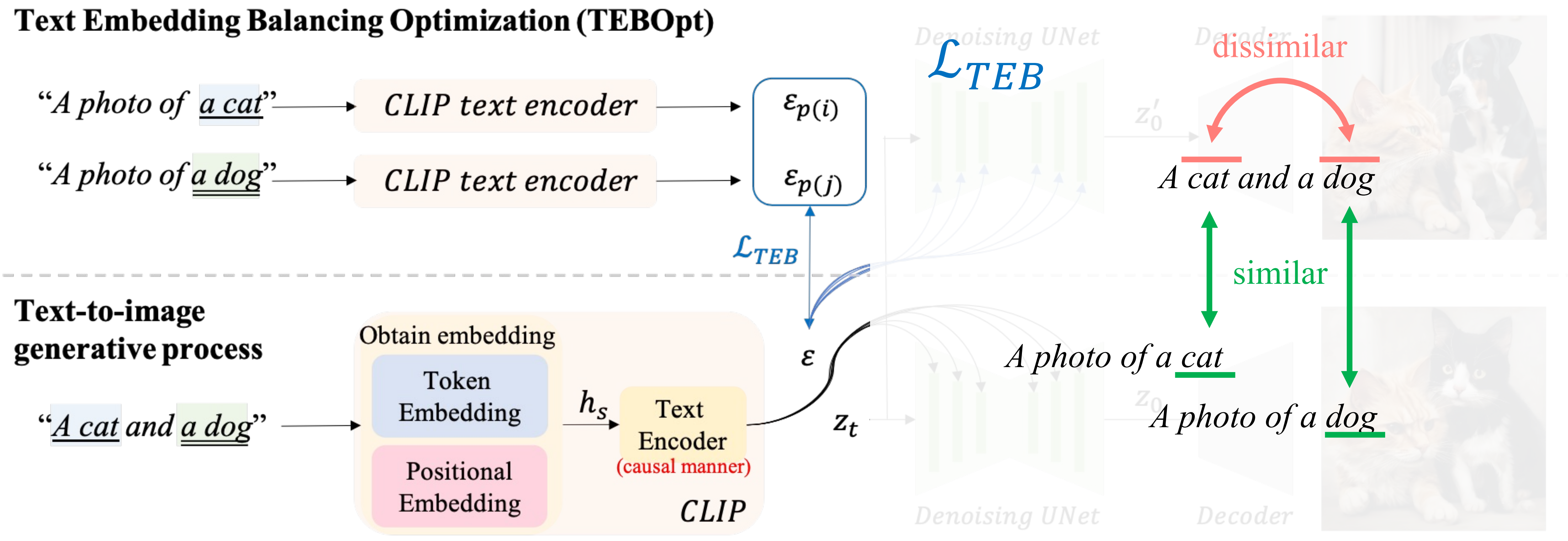
The proposed text embedding optimization method



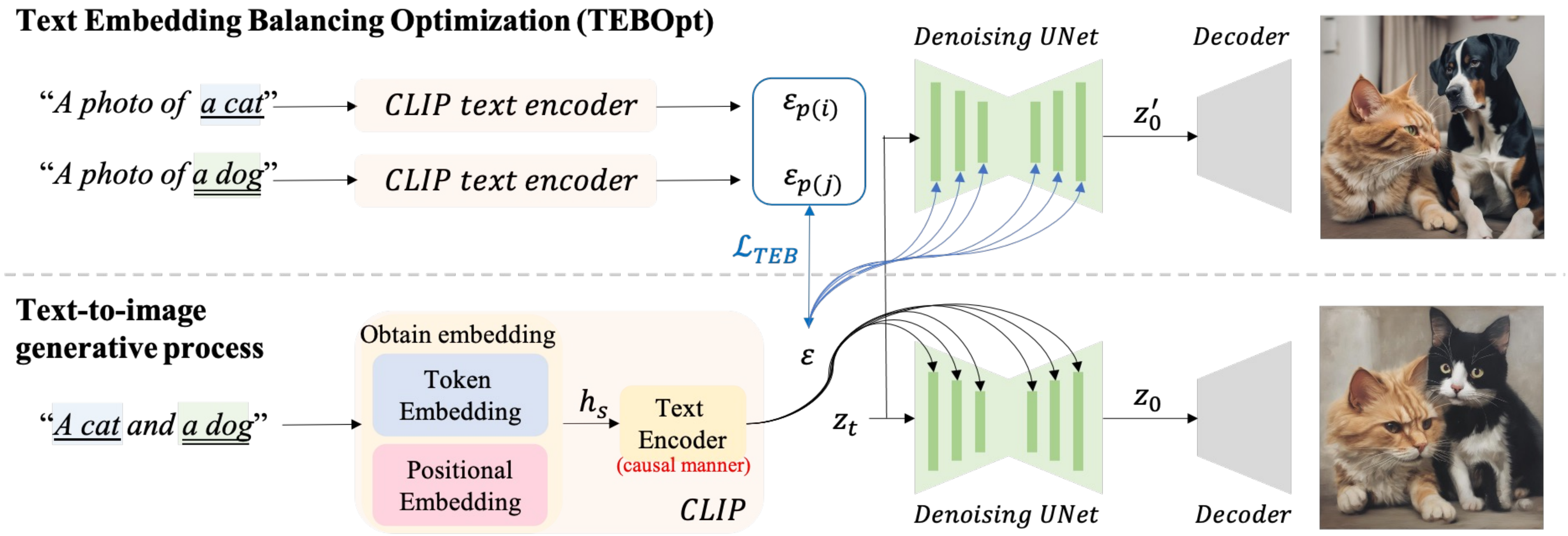
The proposed text embedding optimization method



The proposed text embedding optimization method



The proposed text embedding optimization method



Qualitative Results

SDXL-Turbo Color mixture | Ours | SD 1.4 Car missing | Ours | SD 1.4 Frog missing | Ours

A blue chair and a red cup | *A sheep near a car* | *A bear and a frog*

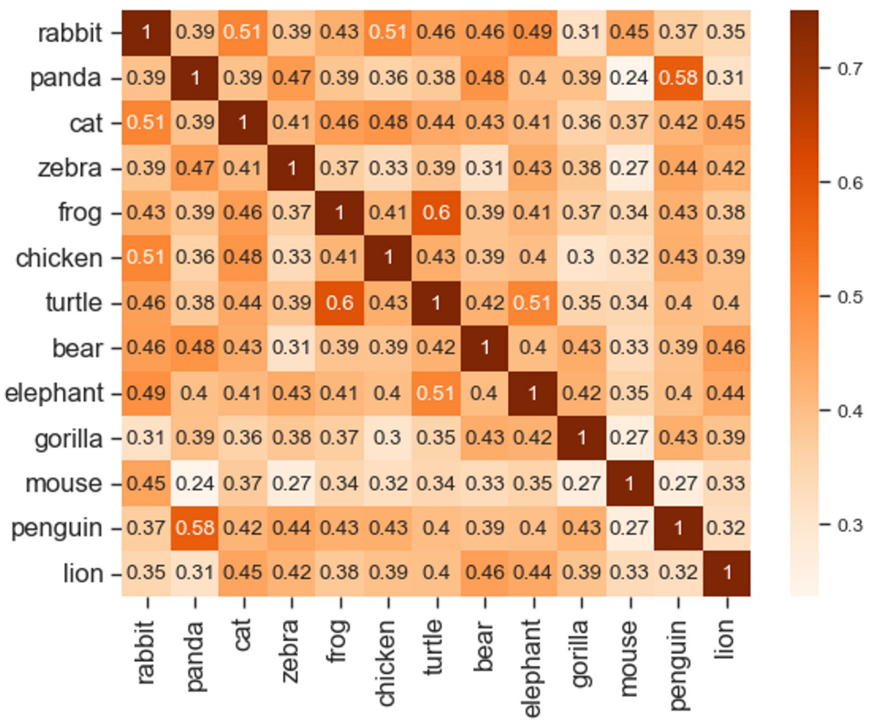
SD 3 Object mixture | Ours | SD 3 Clock missing | Ours | SD 3 Color Mixture Orange Missing | Ours

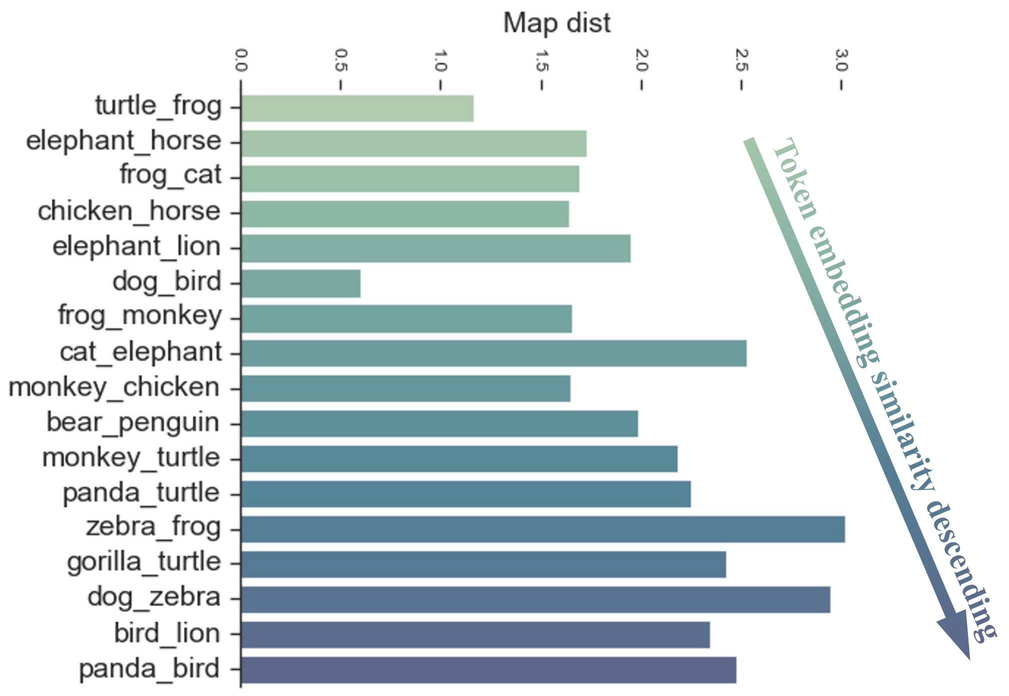
A brown giraffe and a red train | *A brown bench and a green clock* | *A blue bowl and a yellow orange*

Discussion of how the similarity of text embedding affects cross-attention maps' distance

The cosine similarity of text embedding from single word

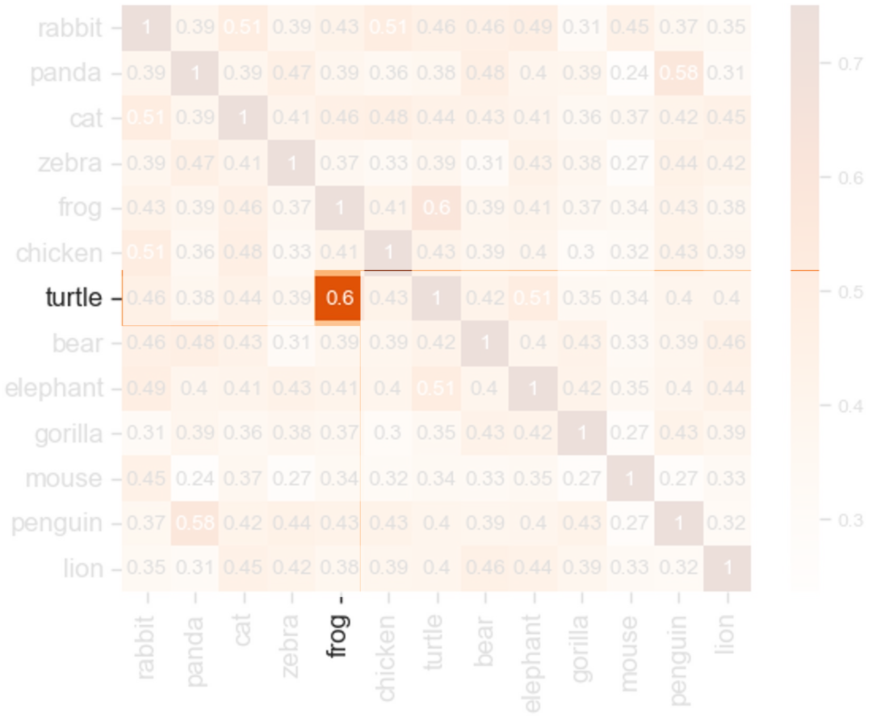
The KL distance of cross-attention maps that are triggered by two words. The data is ordered by their text embedding similarity.





Discussion of how the similarity of text embedding affects cross-attention maps’ distance

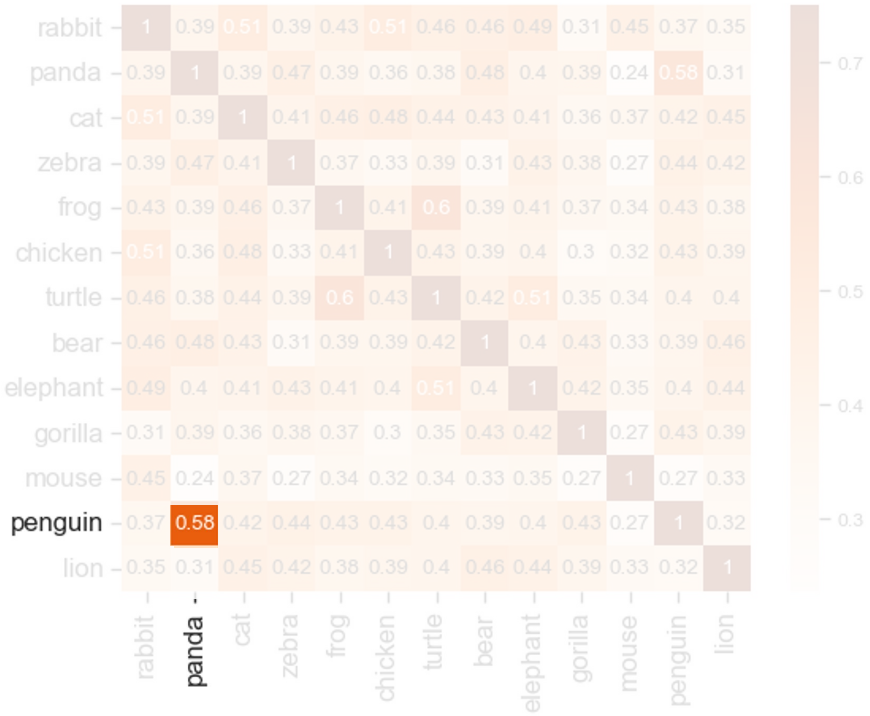
The cosine similarity of text embedding from single word



The KL distance of cross-attention maps that are triggered by two words. The data is ordered by their text embedding similarity.

Discussion of how the similarity of text embedding affects cross-attention maps' distance

The cosine similarity of text embedding from single word

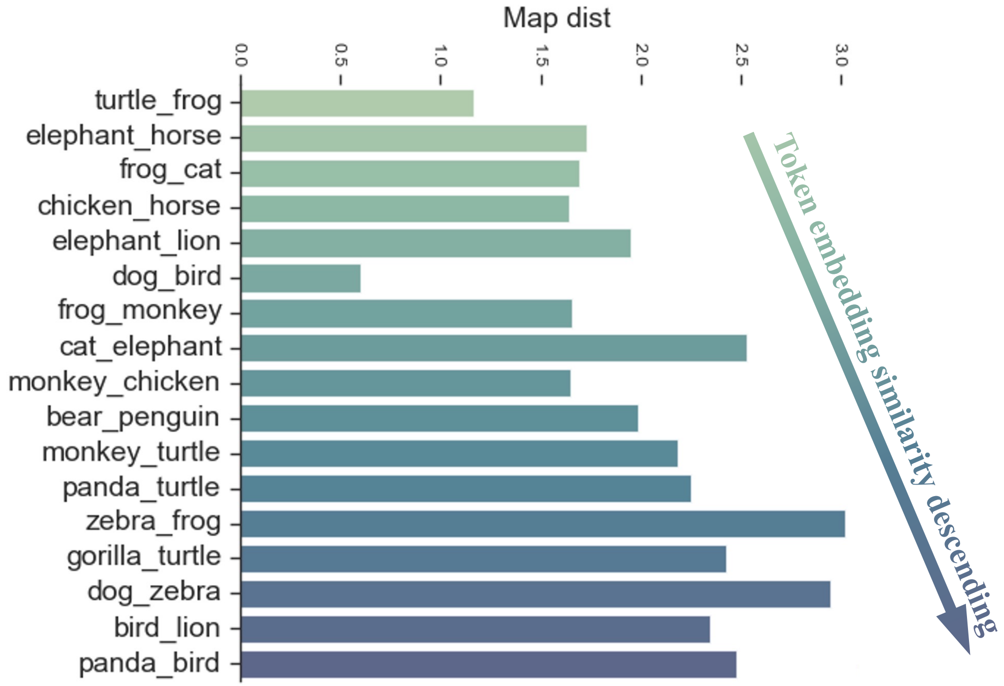


The KL distance of cross-attention maps that are triggered by two words. The data is ordered by their text embedding similarity.

Discussion of how the similarity of text embedding affects cross-attention maps’ distance

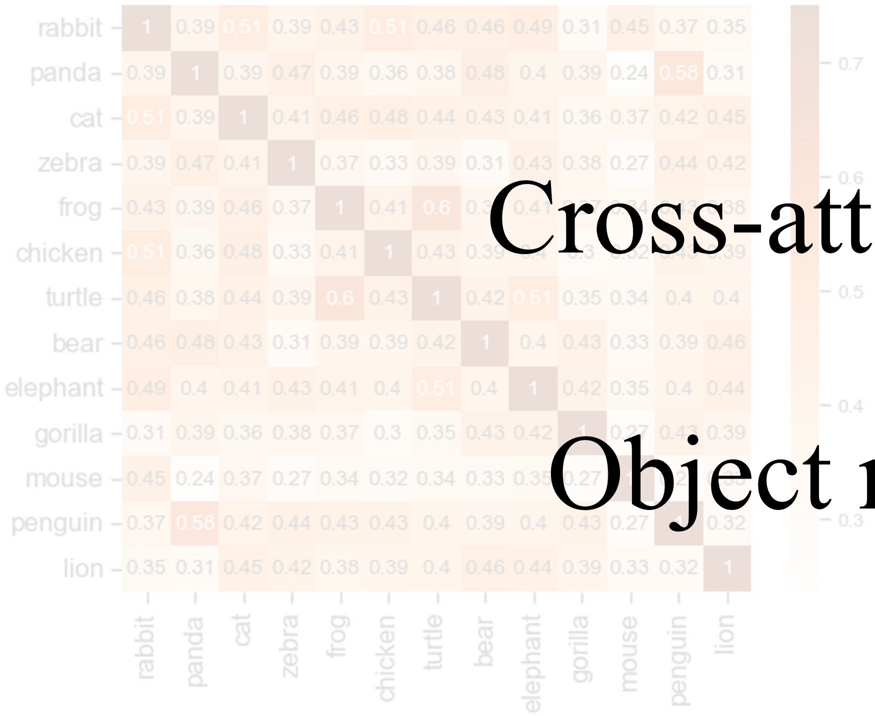
The cosine similarity of text embedding from single word

The KL distance of cross-attention maps that are triggered by two words. The data is ordered by their text embedding similarity.



Discussion of how the similarity of text embedding affects cross-attention maps' distance

The cosine similarity of text embedding from single word



The KL distance of cross-attention maps that are triggered by two words. The data is ordered by their text embedding similarity.

Token similarity

Cross-attention map distance

Object missing or mixture

Conclusion

1. Examining how text embedding contributes to generated images in text-to-image diffusion models
2. Demystifying how the causal manner leads to information bias and loss while contributing to general information
3. Proposing the Text Embedding Balance Optimization solution containing one positive and one negative loss to optimize text embedding for tackling information bias with 125.42% improvement in Stable Diffusion
4. Proposing an evaluation metric to measure information loss. Compared to the CLIP score for evaluating text-image similarity, and the CLIP-BLIP score for evaluating text-text similarity, our evaluation metric provides a concrete number for identifying whether the specified object exists in the generated image.

Thank you for your interest!

Contact: Chieh-Yun Chen (cychenisme@gmail.com)