Skews in the Phenomenon Space Hinder Generalization in Text-to-Image Generation

Yingshan Chang, Yasi Zhang, Zhiyuan Fang, Yingnian Wu, Yonatan Bisk, and Feng Gao

SoTA Text-to-Image Models Struggle at Spatial Relations

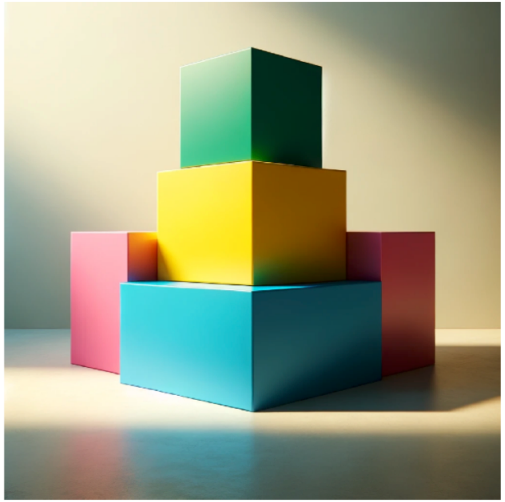
A horse riding an astronaut

(a)

A mouse chasing a cat

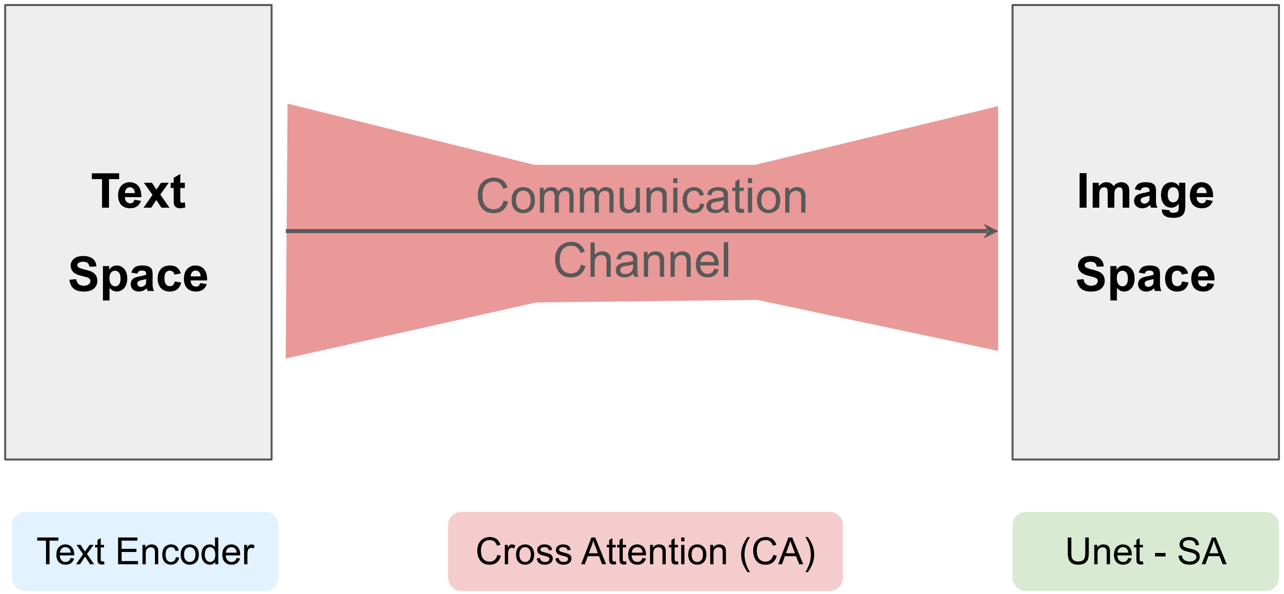
(b)

A pink box is on top of blue box,
which is on top of a yellow box,
which is on top of a green box.



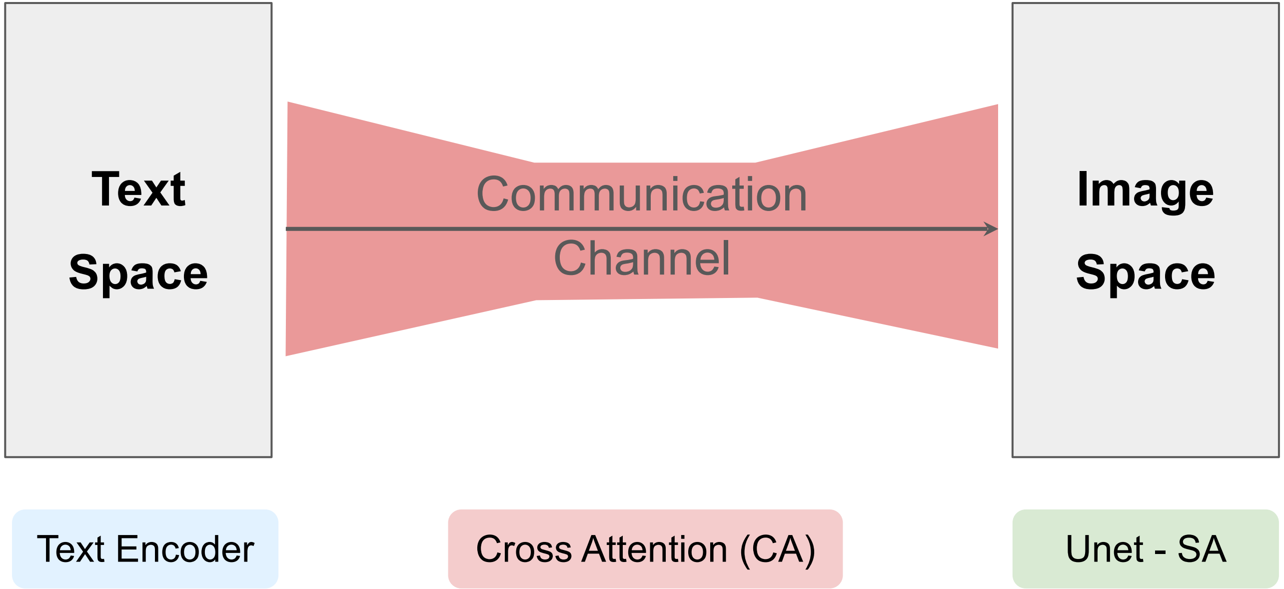
(c)

Error Sources for Faulty Spatial Relations



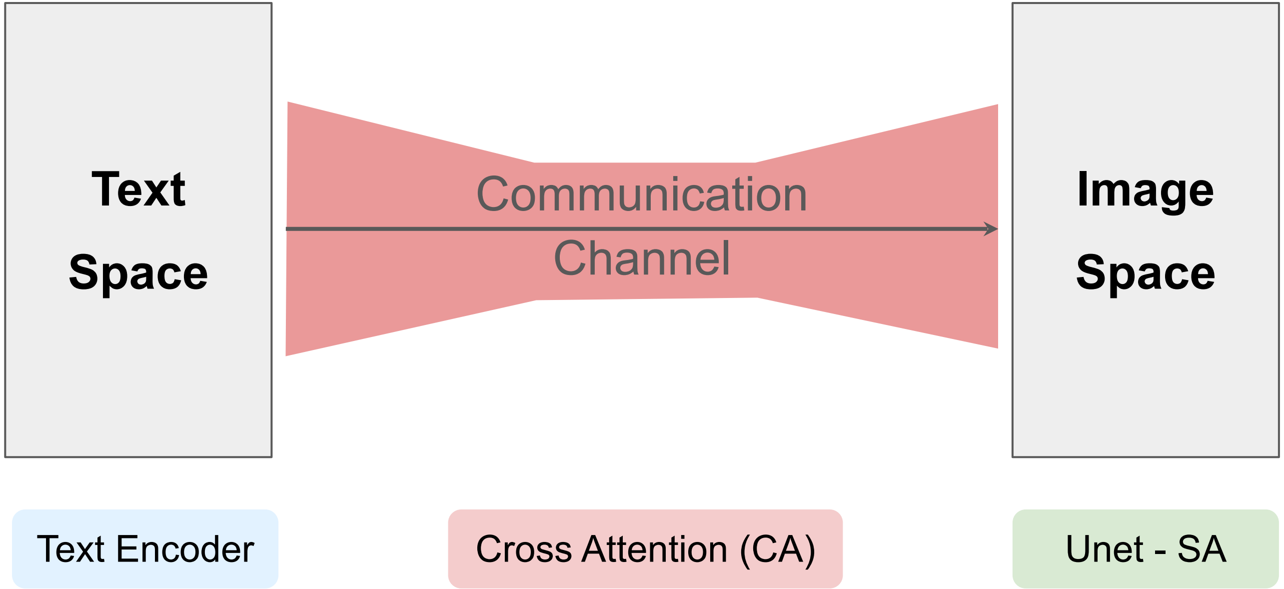
Error Sources for Faulty Spatial Relations

- Text encoder does not correctly encode positions



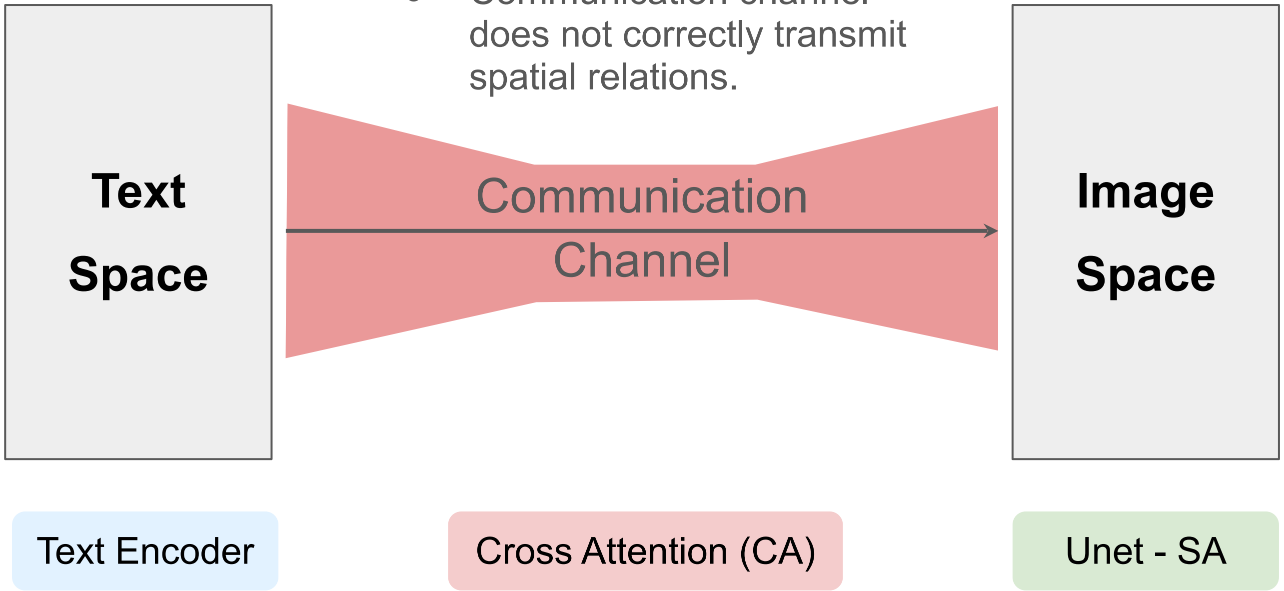
Error Sources for Faulty Spatial Relations

- Text encoder does not correctly encode positions
- Image decoder does not distinguish positions



Error Sources for Faulty Spatial Relations

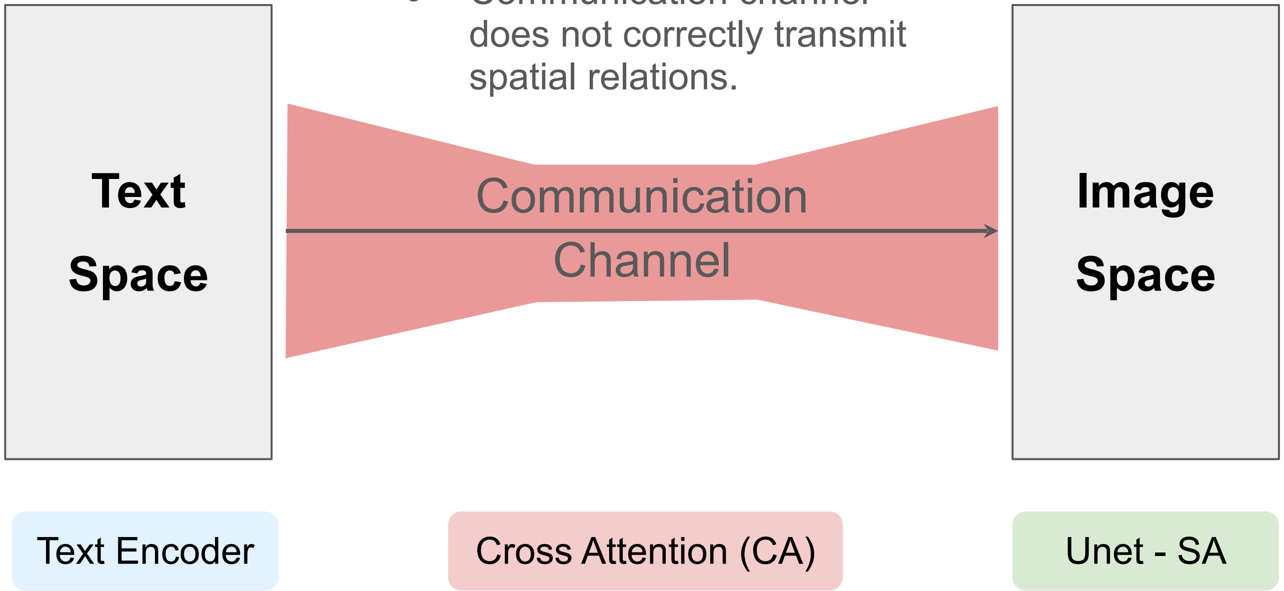
- Text encoder does not correctly encode positions
- Image decoder does not distinguish positions
- Communication channel does not correctly transmit spatial relations.



Error Sources for Faulty Spatial Relations

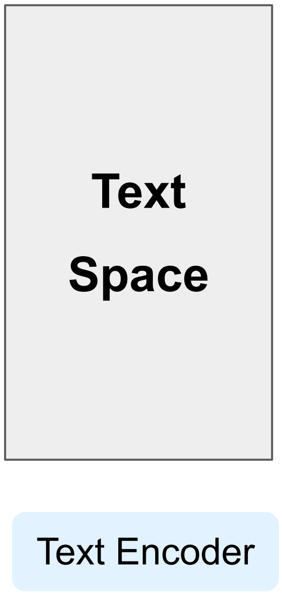
Isolate the error source with experiments

- Text encoder does not correctly encode positions
- Communication channel does not correctly transmit spatial relations.
- Image decoder does not distinguish positions



Error Source 1: Are positions correctly encoded from text?

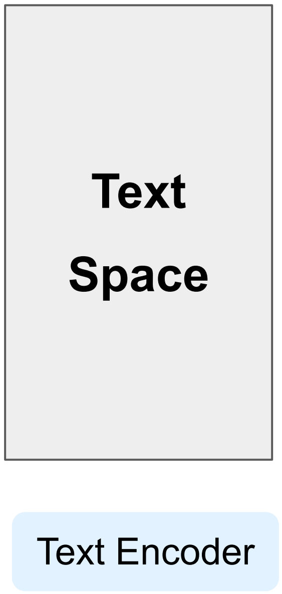
- Text encoder does not correctly encode positions

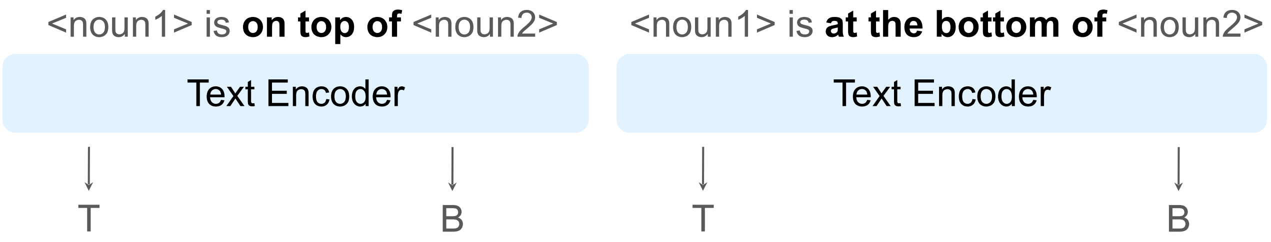


Error Source 1: Are positions correctly encoded from text?

Experiment: probe position information from token encodings

- Text encoder does not correctly encode positions



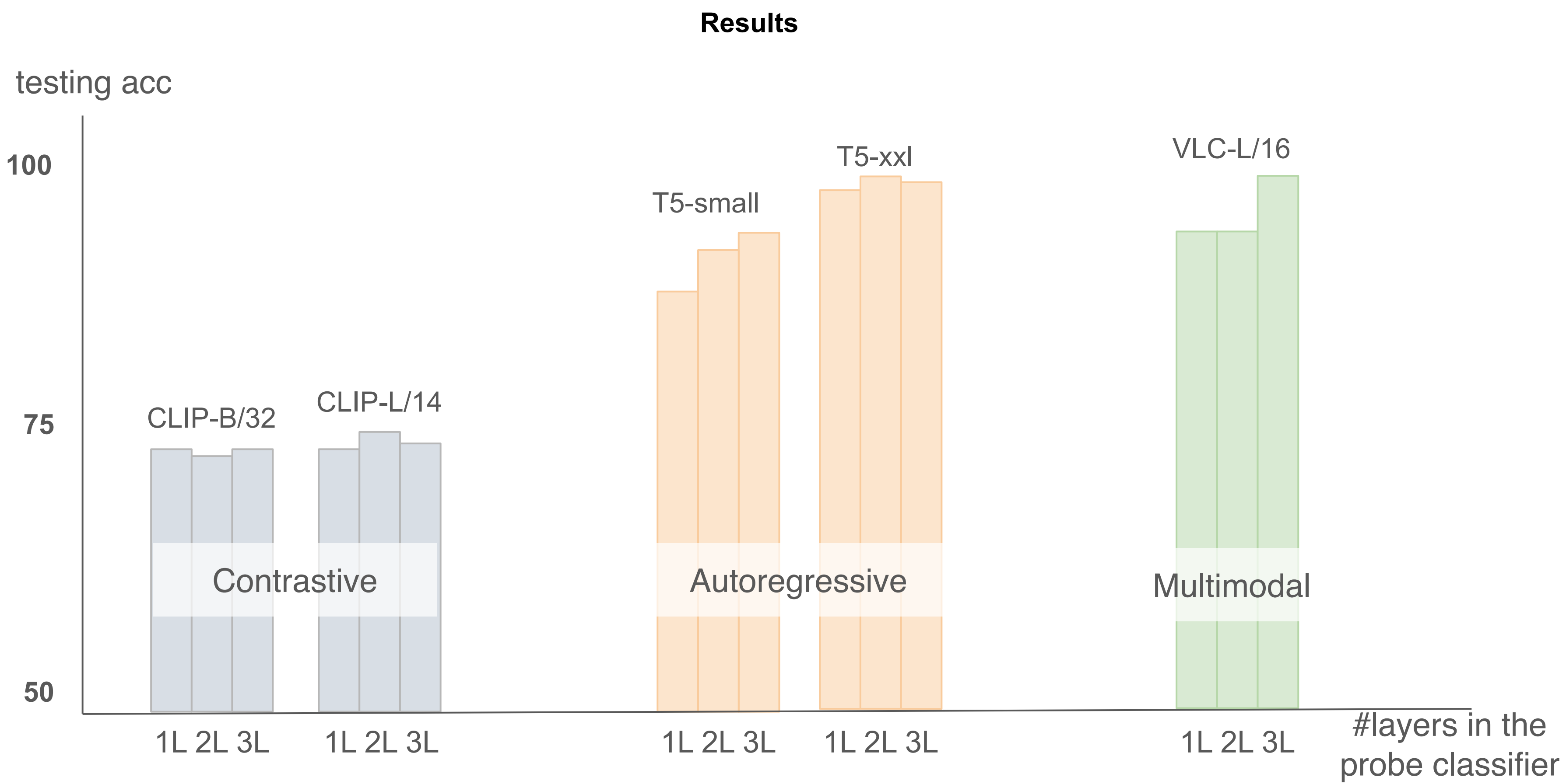


Training: noun1, noun2 randomly sampled from S_{train} = {English nouns}.

Testing: noun1, noun2 randomly sampled from S_{test};

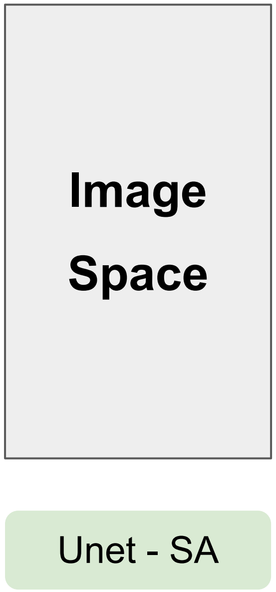
S_{test} and S_{train} don't overlap.

Error Source 1: Are positions correctly encoded from text?



Error Source 2: Is position info available in image decoders?

- Image decoder does not distinguish positions (e.g. being invariant to positions)



Error Source 2: Is position info available in image decoders?

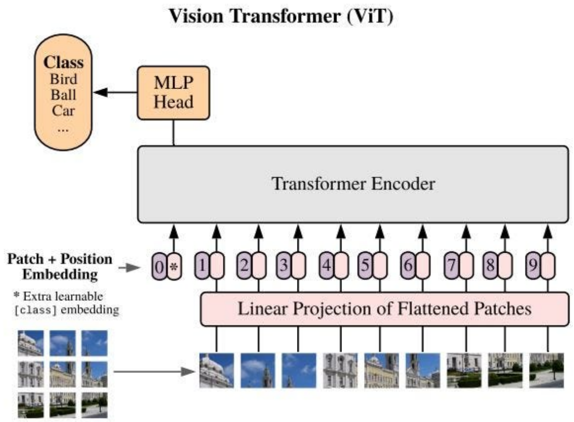
Experiment: ablate position-embeddings from image decoders

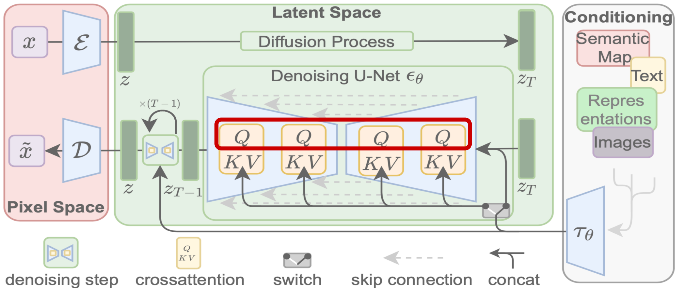
- Image decoder does not distinguish positions (e.g. being invariant to positions)

Image

Space

Unet - SA

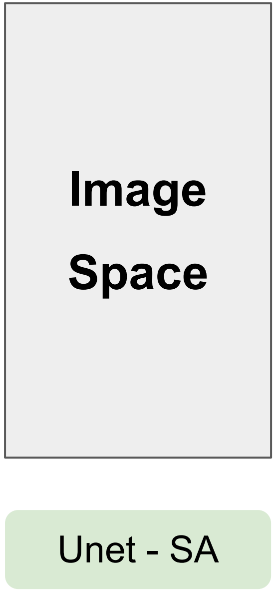




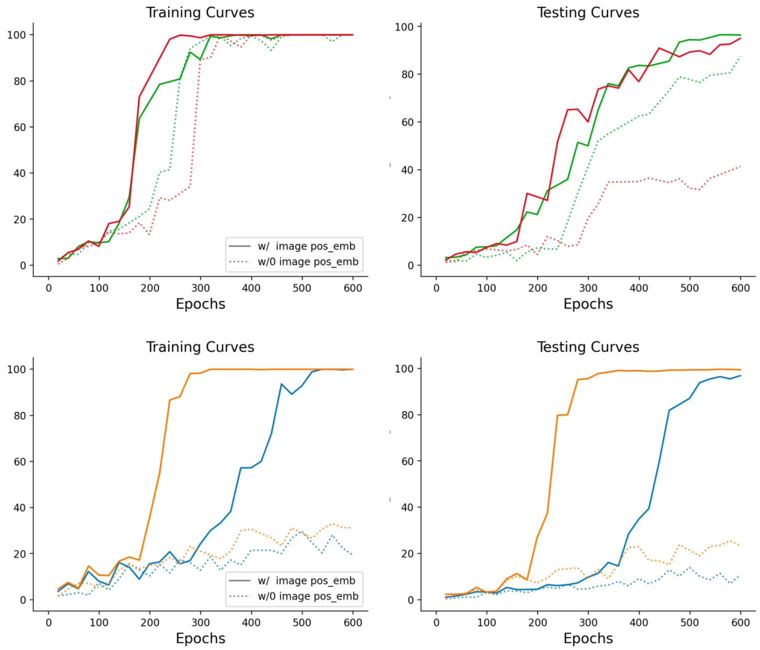
Stable Diffusion: No image positional embeddings by default!

Error Source 2: Is position info available in image decoders?

- Image decoder does not distinguish positions (e.g. being invariant to positions)



Results



Models w/o image positional embeddings exhibit both **slower convergence in training** and **worse performance in testing**.

Error Source 3: Faulty communication of spatial relations

- Text encoder does not correctly encode positions

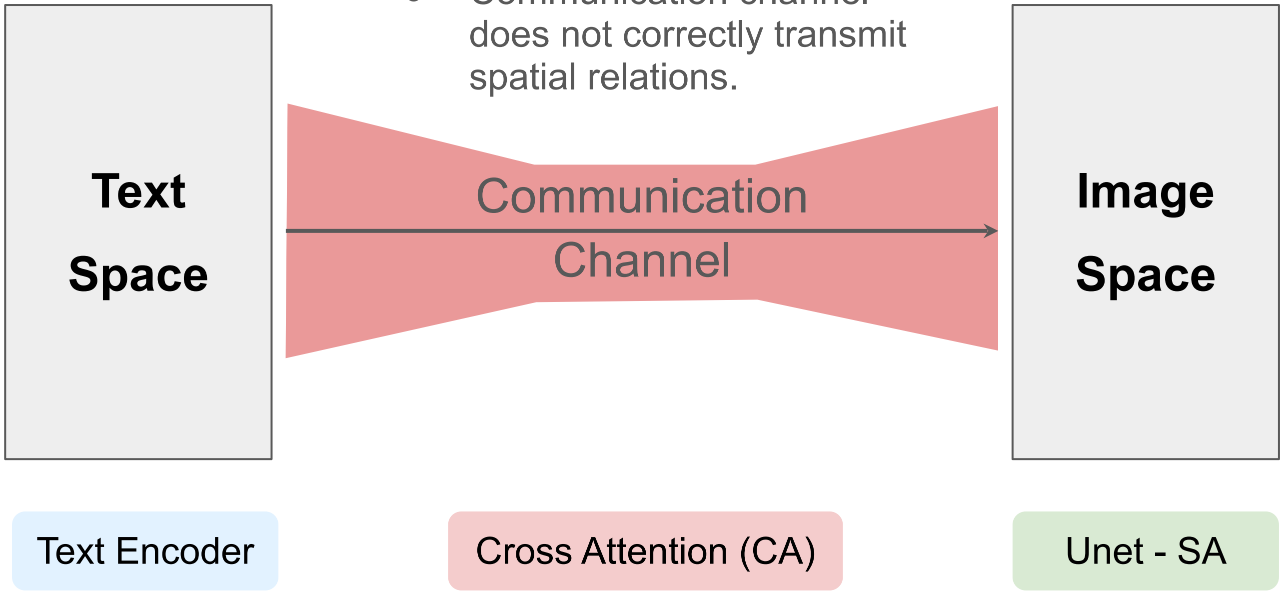
Prefer T5, VLC, avoid CLIP

- Communication channel does not correctly transmit spatial relations.

?

- Image decoder does not distinguish positions (e.g. being invariant to positions)

Image Positional Embeddings



Why is the communication of spatial relations hard?

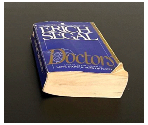
- A relation does not take any perceptible form. — Abstractness
- A relation can only be observed with concrete objects, but it should not be permanently associated with those objects. — Disentanglement
- A relation can be associated with “unseen” objects after learning. — Composition

Formalization

When a message specifies spatial relation between objects, what a formal structure does it entail?

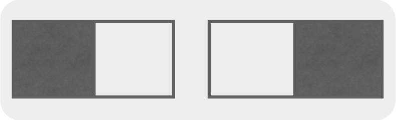
Role-filler bindings

Fillers: concrete values, e.g. objects



cap book

Roles: abstract positions



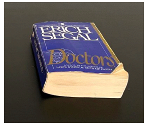
subj obj

Formalization

When a message specifies spatial relation between objects, what a formal structure does it entail?

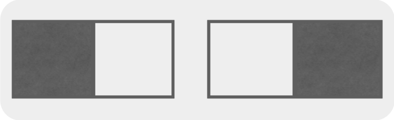
Role-filler bindings

Fillers: concrete values, e.g. objects



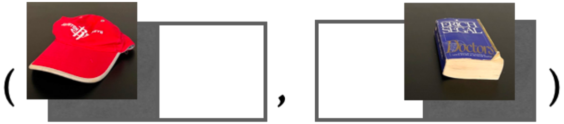
cap book

Roles: abstract positions



subj obj

Binding: put fillers into the roles



(book-subj, cap-obj)

Each image/caption corresponds to a role-filler binding.

A book right of a cap

Error Source 3: Faulty communication of spatial relations

When a message specifies spatial relation between objects, what a formal structure does it entail?

Role-filler bindings

Possible cause for a faulty communication channel:

Training data only covers a **skewed** space of this formal structure.

Phenomenological space

?

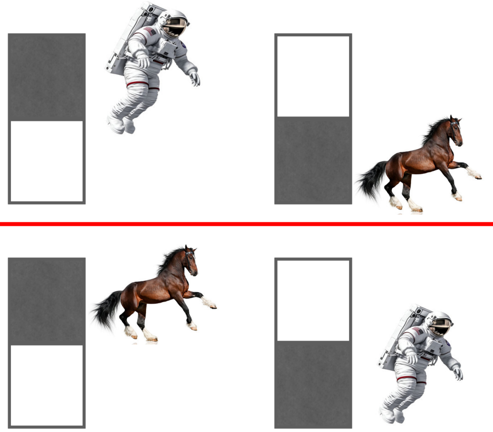
- Communication channel does not correctly transmit spatial relations.

Cross Attention (CA)

Error Source 3: Faulty communication of spatial relations

(astronaut-subj, horse-obj)

(horse-subj, astronaut-obj)

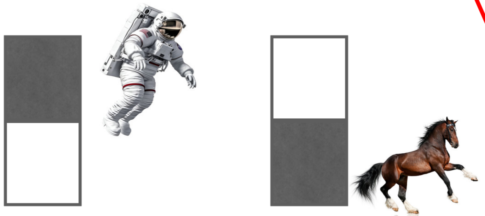


Missing

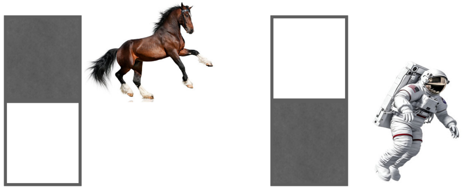
An astronaut riding a horse

Error Source 3: Faulty communication of spatial relations

(astronaut-subj, horse-obj)



(horse-subj, astronaut-obj)



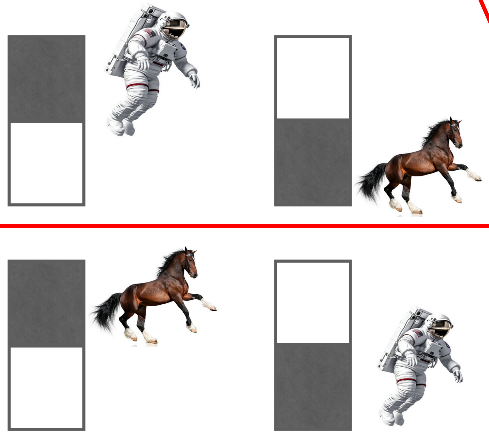
Missing

An astronaut riding a horse

Error Source 3: ~~Faulty communication of spatial relations~~

(astronaut-subj, horse-obj)

(horse-subj, astronaut-obj)



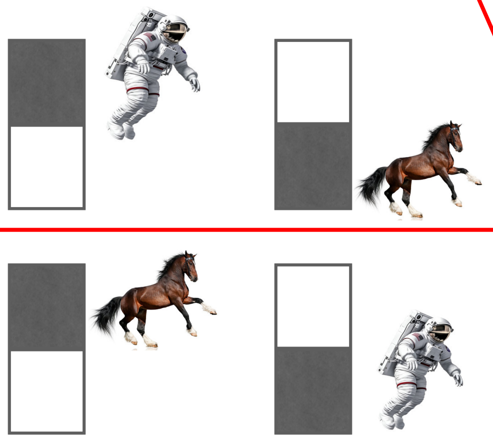
Missing

An astronaut riding a horse

Error Source 3: Skew in the phenomenological space!

(astronaut-subj, horse-obj)

(horse-subj, astronaut-obj)

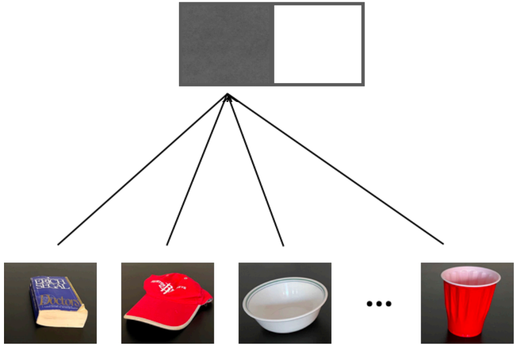


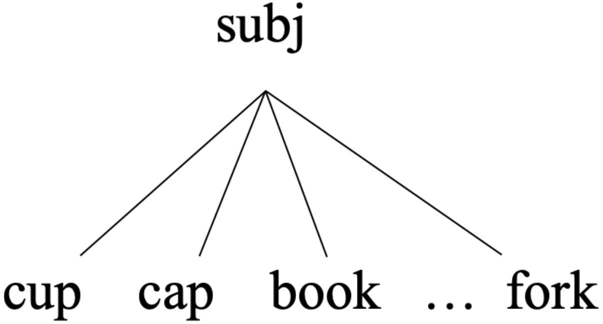
Missing

An astronaut riding a horse

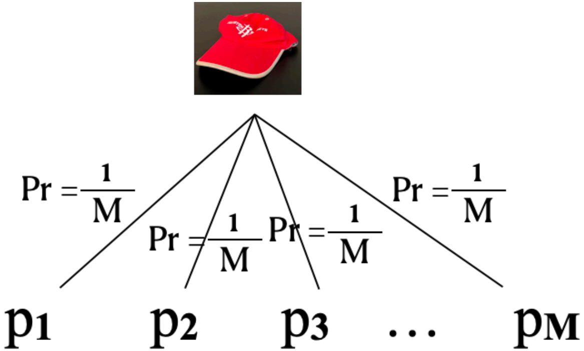
Two Metrics to Quantify Skew

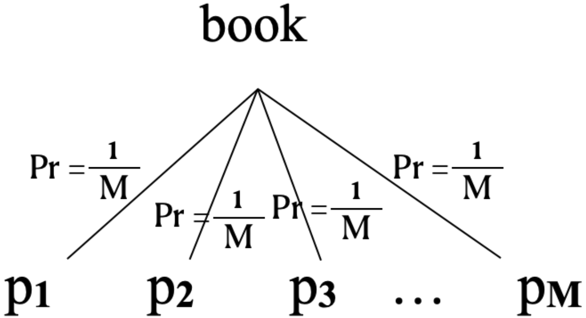
Completeness





Balance

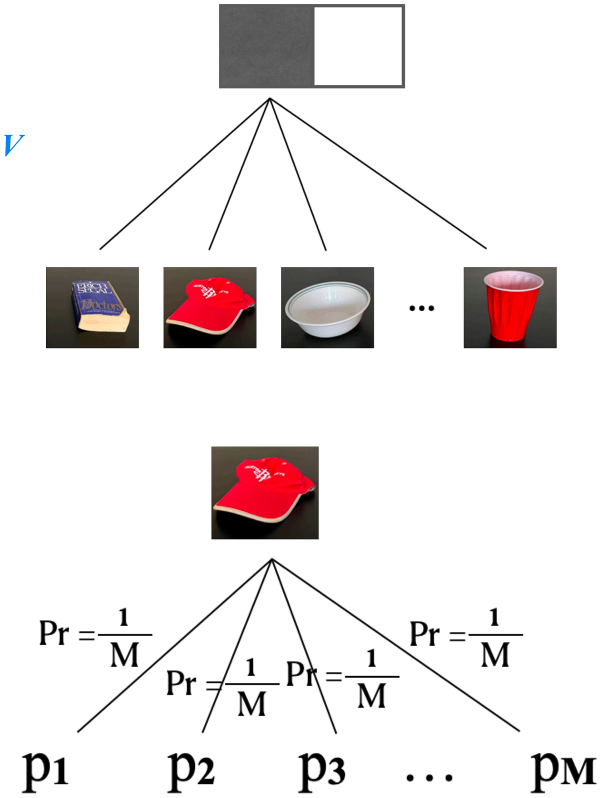


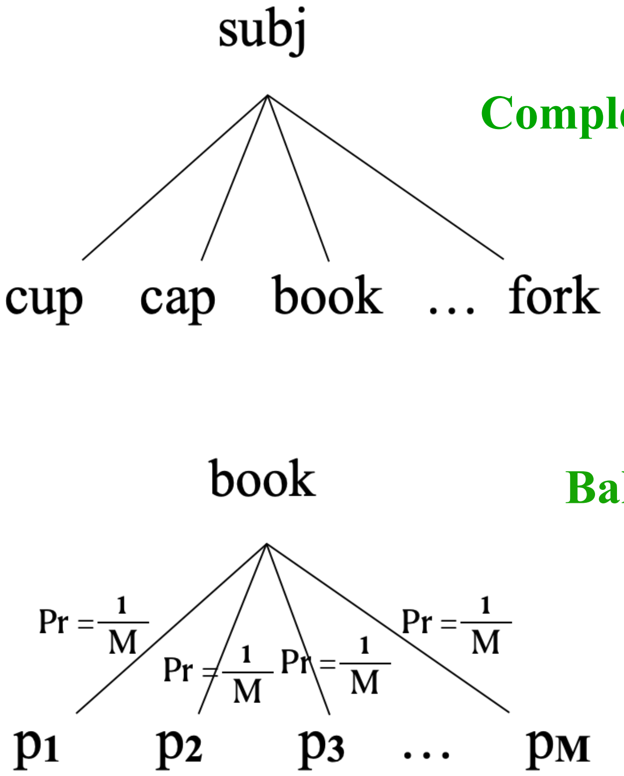


Two Metrics to Quantify Skew

Completeness$_V$

Balance$_V$





Completeness$_L$

Balance$_L$

Skew Hinders Learning Generalizable Spatial Relations

Experimental Design

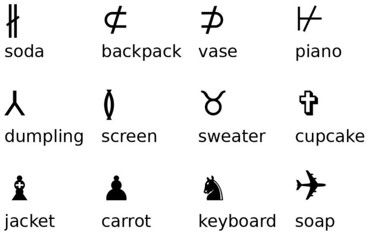
Hypothesis: Completeness$_L$, Balance$_L$, Completeness$_V$, Balance$_V$ Generalization

Visual fillers:

Synthetic icons

Linguistic fillers:

English nouns



Skew Hinders Learning Generalizable Spatial Relations

Experimental Design

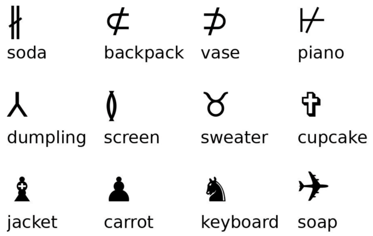
Hypothesis: Completeness_L, Balance_L, Completeness_V, Balance_V Generalization

Visual fillers:

Synthetic icons

Linguistic fillers:

English nouns



Roles:

Top

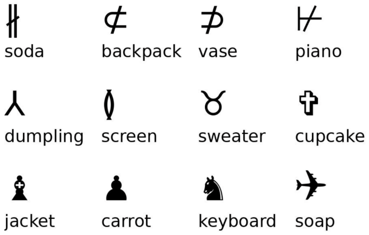
Bottom

Skew Hinders Learning Generalizable Spatial Relations

Experimental Design

Hypothesis: Completeness_L, Balance_L, Completeness_V, Balance_V ↓ Generalization ↓

Visual fillers:
Synthetic icons
Linguistic fillers:
English nouns



Create training sets with varied completeness (CPL) and balance (BLC) scores.

Roles:
Top
Bottom

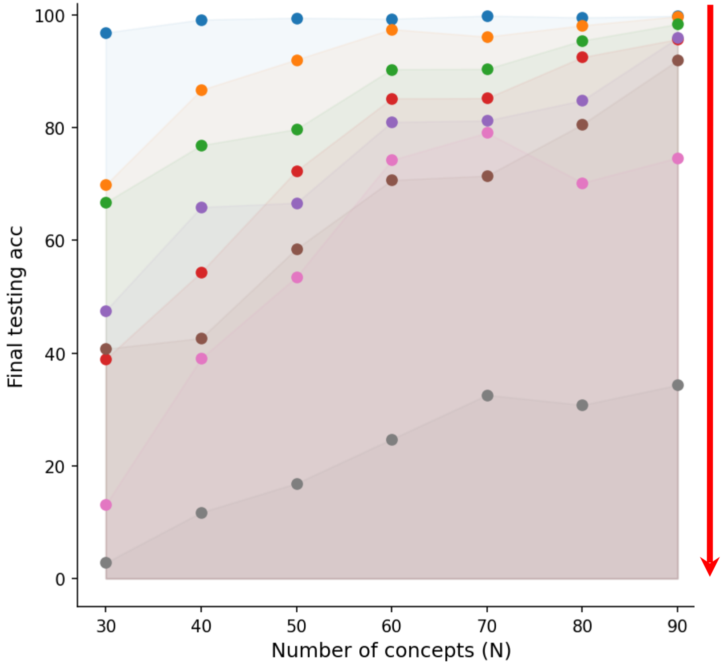
Shaded: role-filler bindings covered by **training** data
Empty: unseen role-filler bindings held-out for **testing**

Metric: accuracy (1 if both objects and the relation is correct, otherwise 0)

Skew Hinders Learning Generalizable Spatial Relations

Results

	CPL(r1)	CPL(r2)	BLC
	100	100	100
	100	100	84
	100	100	79
	100	100	71
	84	84	79
	75	75	68
	100	50	68
	50	50	0



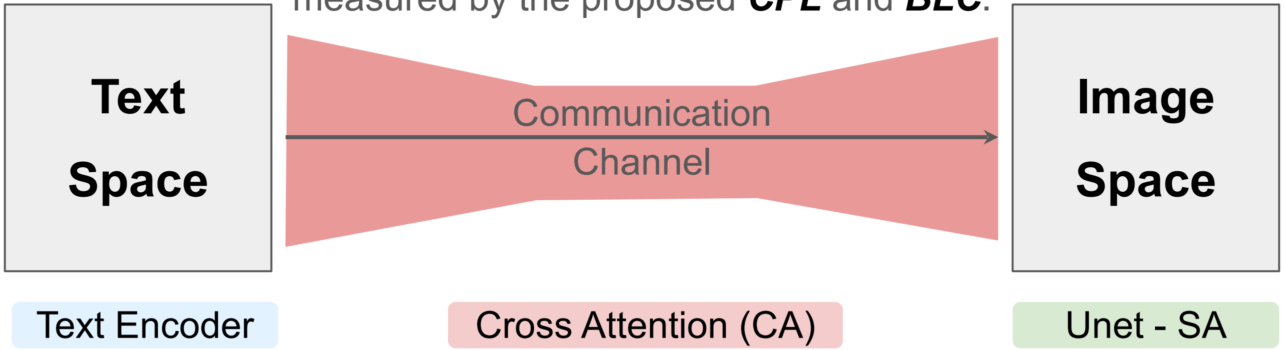
Incomplete or unbalanced data greatly harms generalization.

That's Everything!

Thank you!

Why current text-to-image models are prone to faulty spatial relations?

Reduce ***phenomenological skew*** ——
measured by the proposed ***CPL*** and ***BLC***.



Autoregressive & ***multimodal*** models should be preferred over ***contrastively*** pre-trained ones.

Make sure the image SA layers take ***image positional embeddings*** as input.

New ***dataset/augmentation*** with better CPL and BLC
New ***architecture*** that generalizes even when trained under skewed data source.