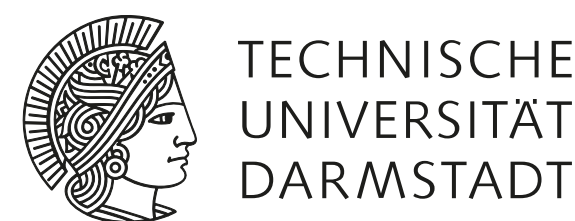


When do Diffusion Models learn to Generate Multiple Objects?

Yujin Jeong, Arnas Uselis, Iro Laina, Seong Joon Oh, Anna Rohrbach

TU Darmstadt & hessian.AI, Tübingen University, University of Oxford, KAIST



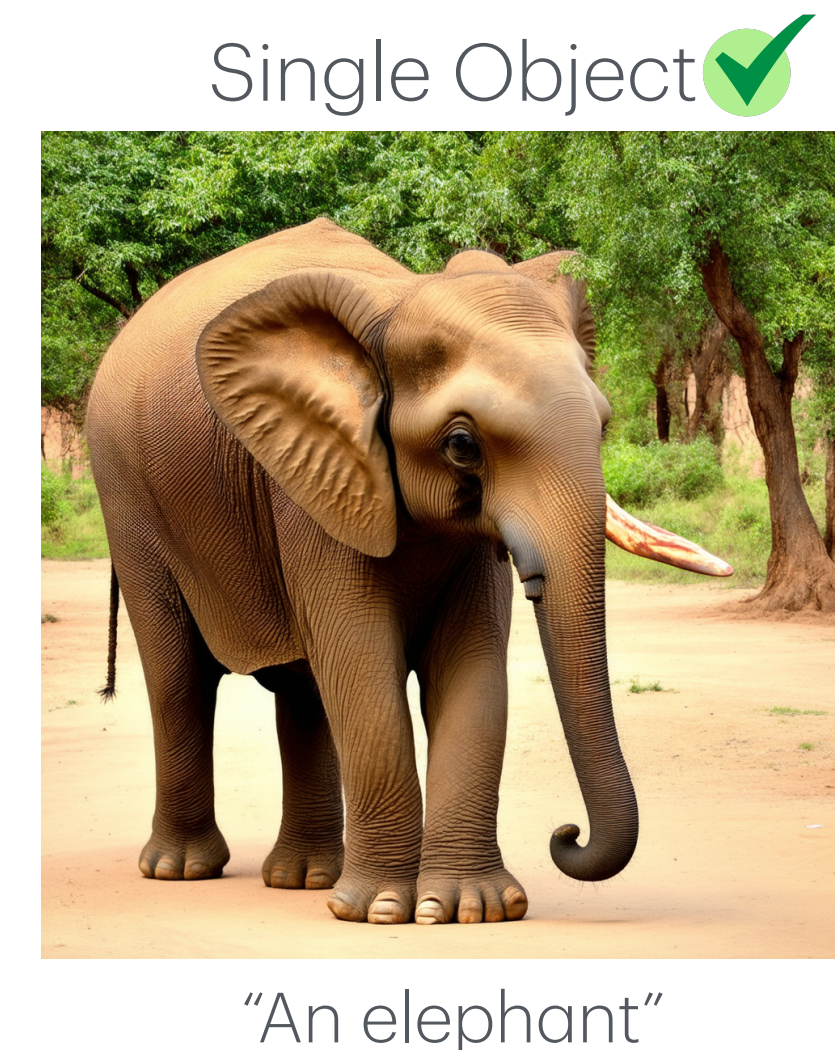
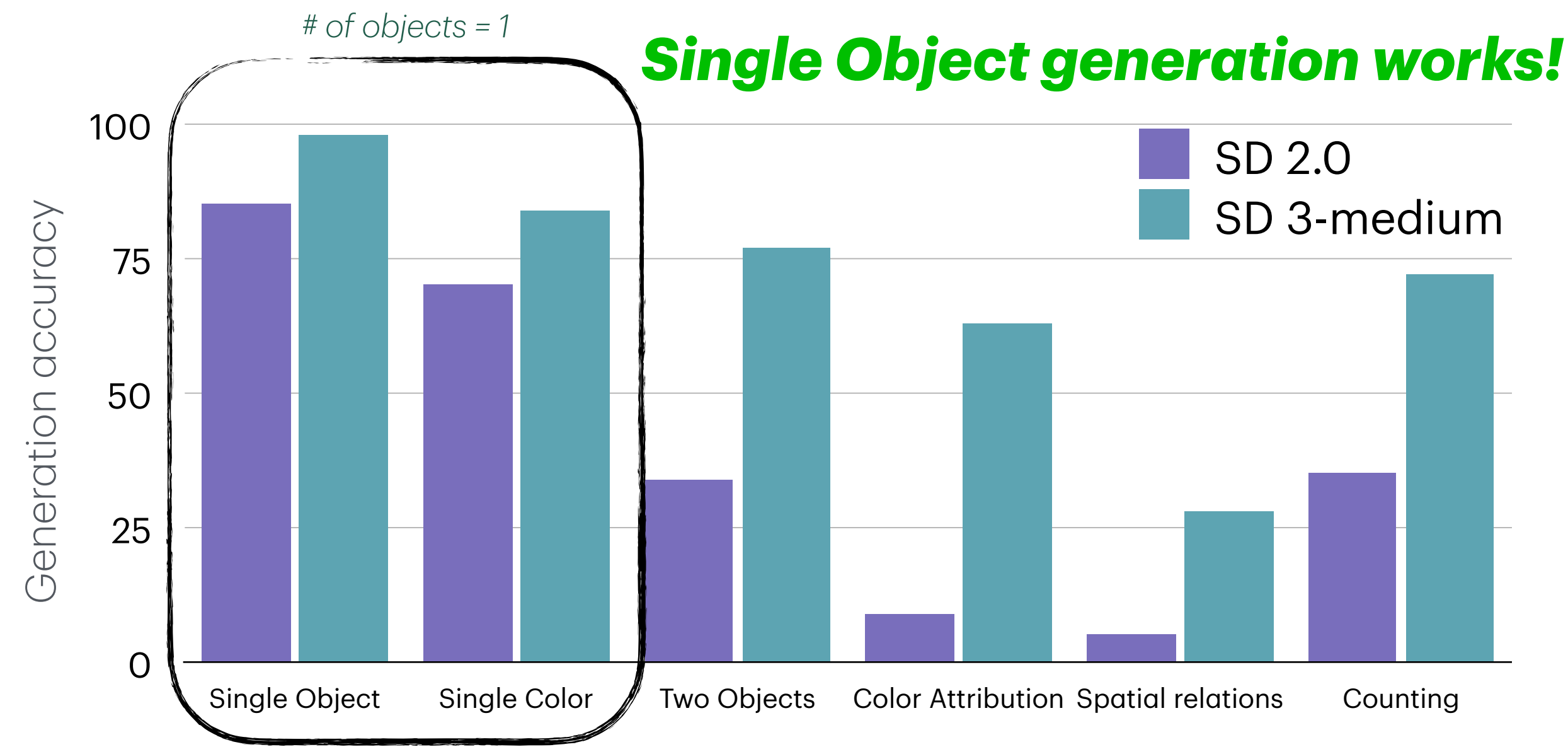
Motivation

Text-to-Image Diffusion models struggle with multi-object generation

Motivation

Text-to-Image Diffusion models struggle with multi-object generation

- GenEval [1] generation accuracy with Stable Diffusion 2 [2] and Stable Diffusion 3 [3] shows strong single-object performance.
- Qualitative examples from Stable Diffusion 3 [3] also show successful single-object generation.

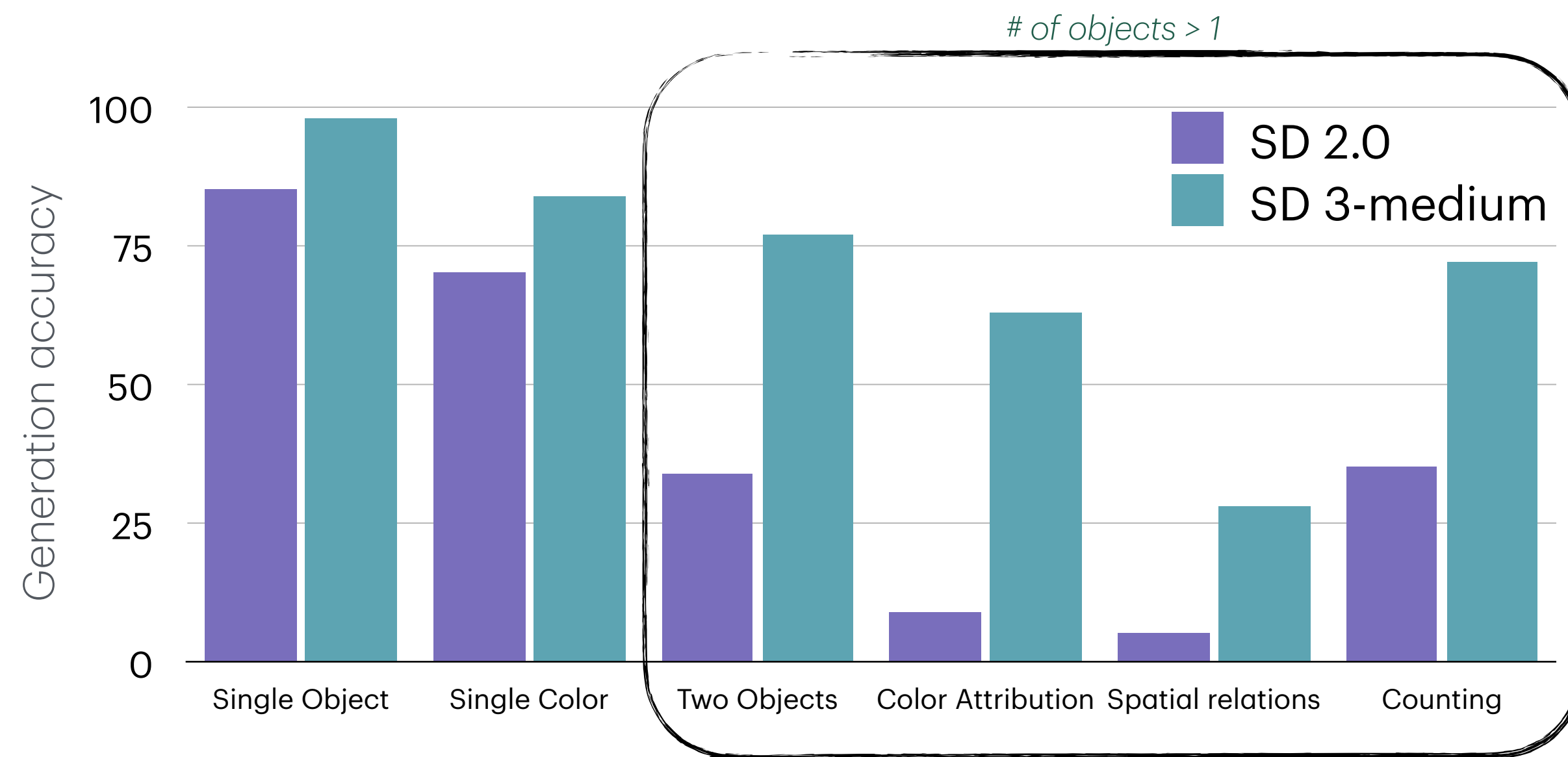


Motivation

Text-to-Image Diffusion models struggle with multi-object generation

- A larger drop on multi-object generation tasks

Multi-object doesn't work!



Two Objects ❌



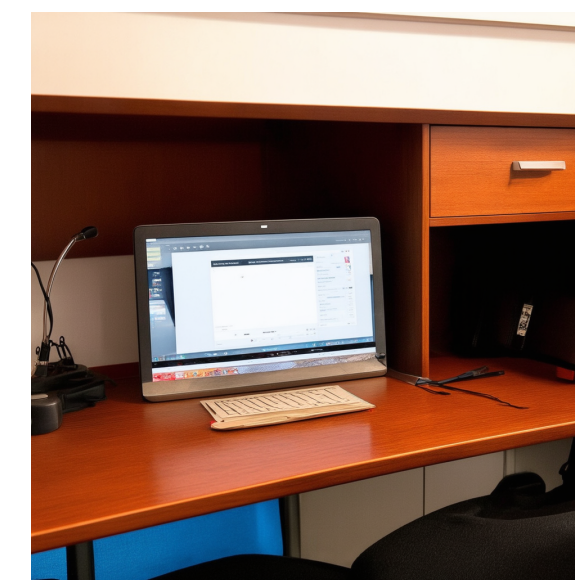
"A zebra and a giraffe"

Color Attribution ❌



"Black cup and yellow tv"

Spatial relations ❌



"A train under the desk"

Counting ❌



"Six skateboards"

Why can't the model generate multi-object?

Why can't the model generate multi-object?

Let's look at the training data!

Why can't the model generate multi-object?

Let's look at the training data!

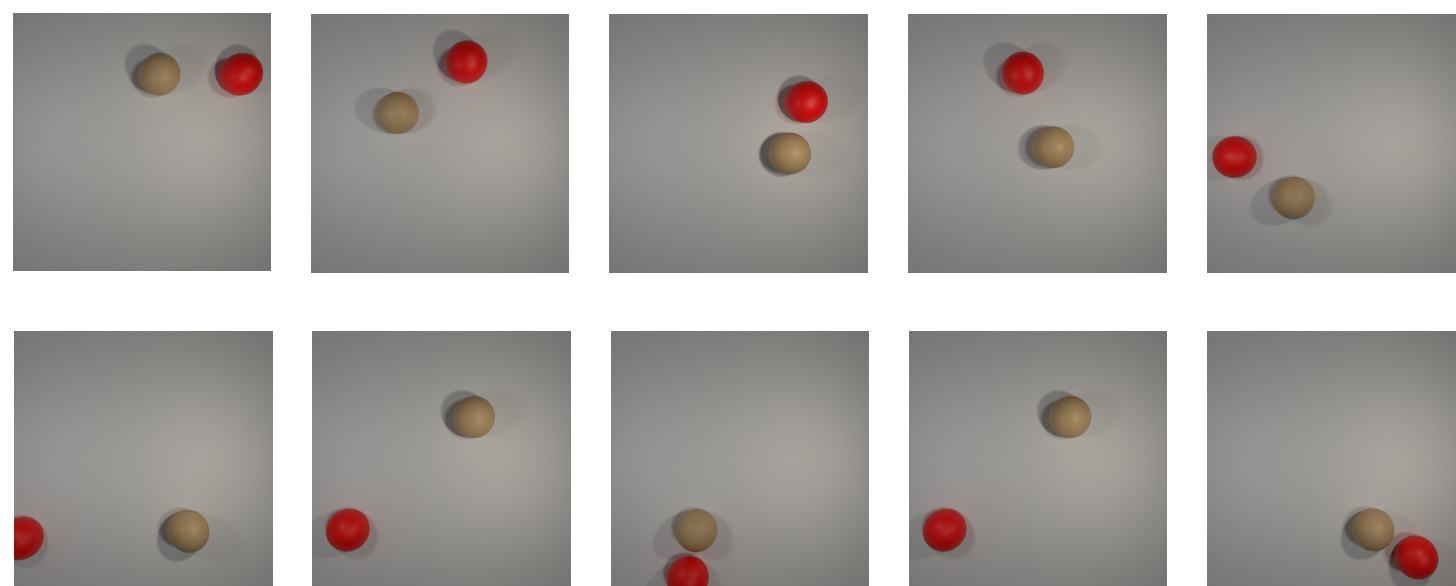
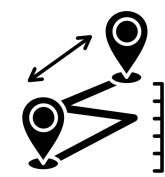
Q. *How reliably can the models generate multi-object compositions under **imperfect training data distributions?** (e.g., imbalanced data, unseen compositions)*

MOSAIC

Multi-**O**bject **S**patial relations, **A**tt**r**ibution, **C**ounting

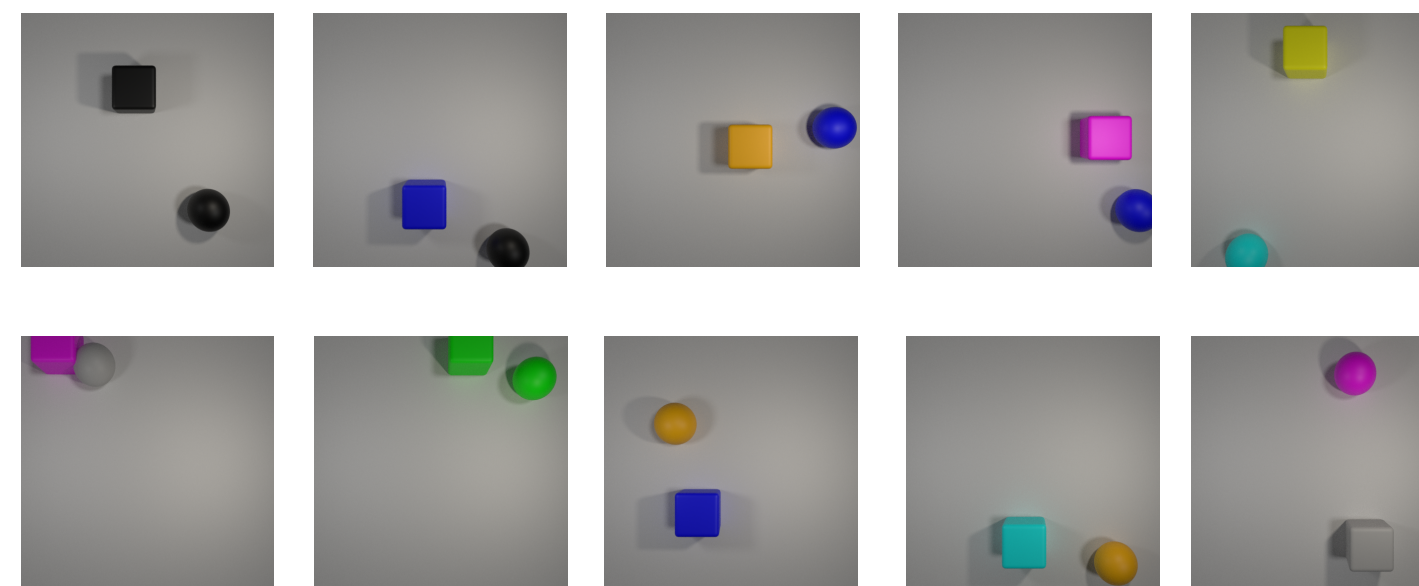
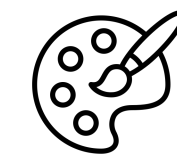
- Built using COMFORT [1] simulator.
- We define three subsets. Each subset isolates one factor of variation and assigns a corresponding class to each image.

Spatial relations



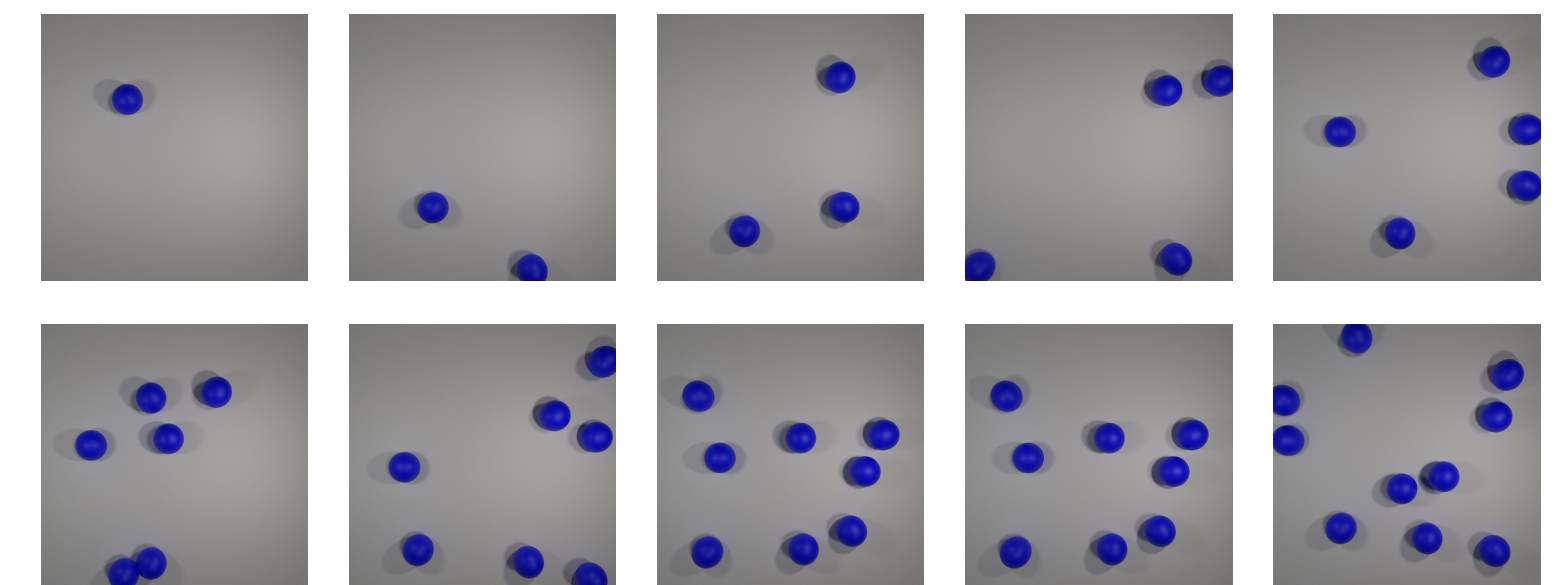
Relative angle between the two objects

Attribution



color combination of different object types
(sphere and cube)

Counting

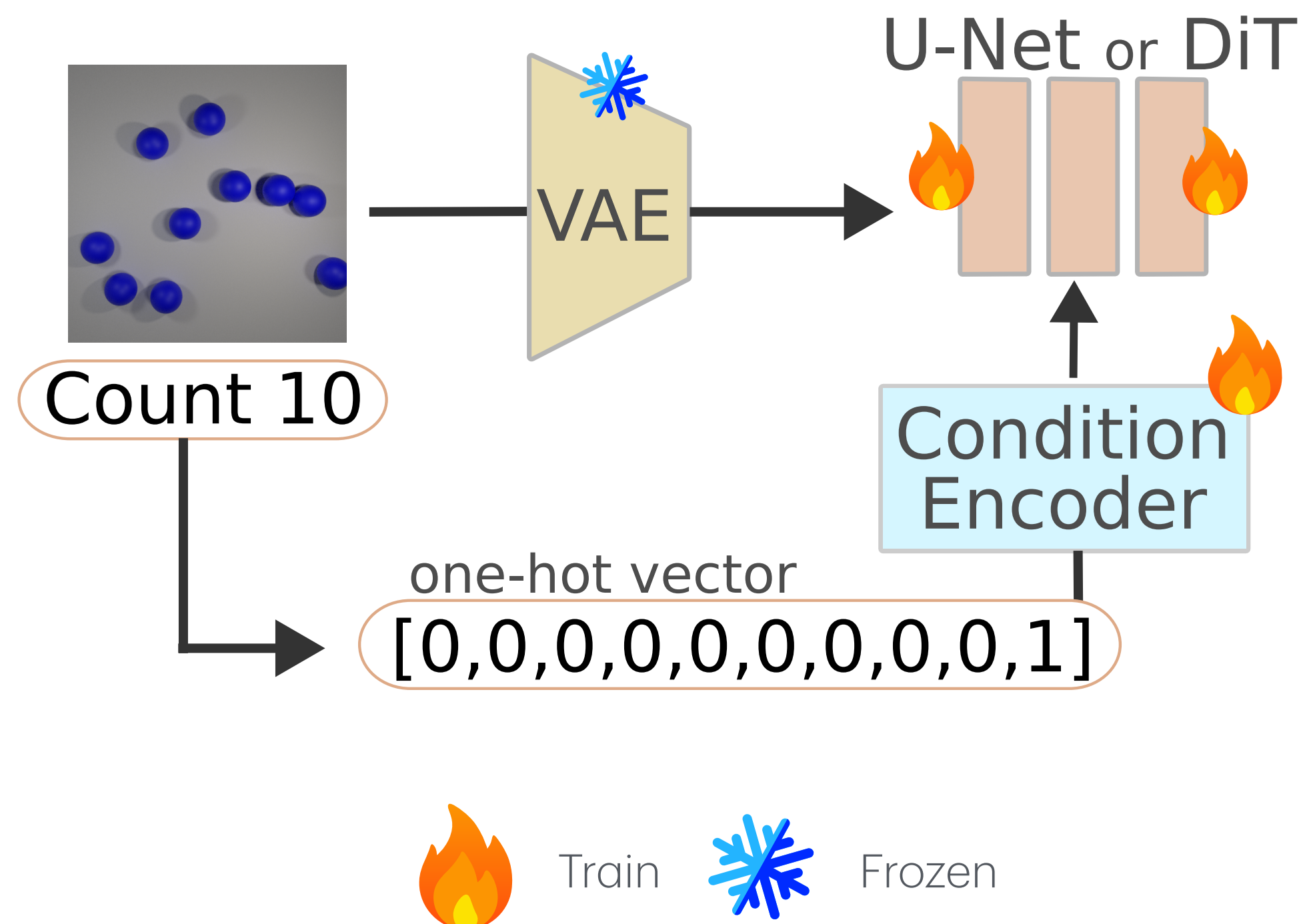


the number of objects in the scene

Experimental Settings

Training and Evaluation Pipeline

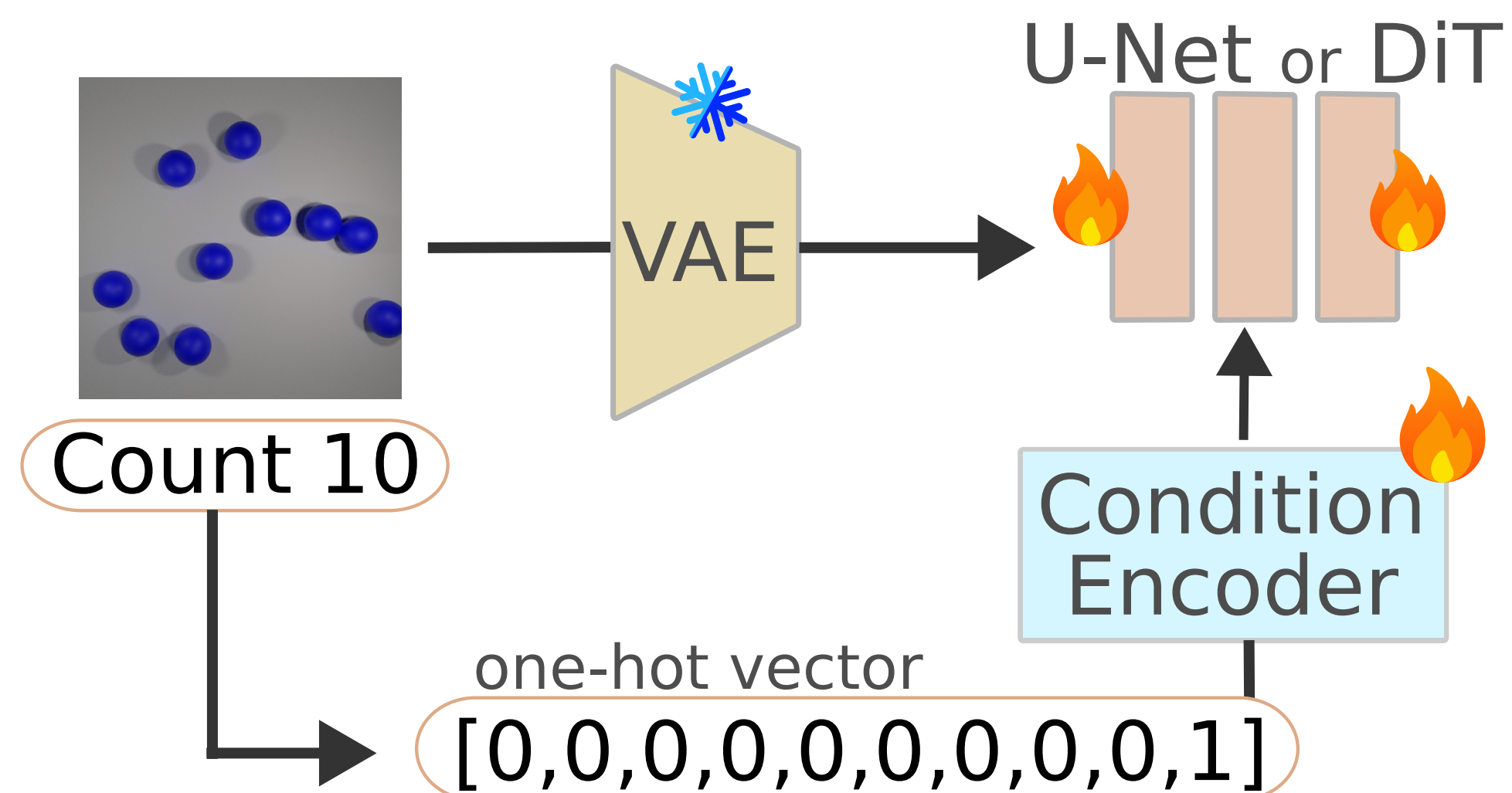
- **Training stage:** Train conditional diffusion models (U-Net, DiT)



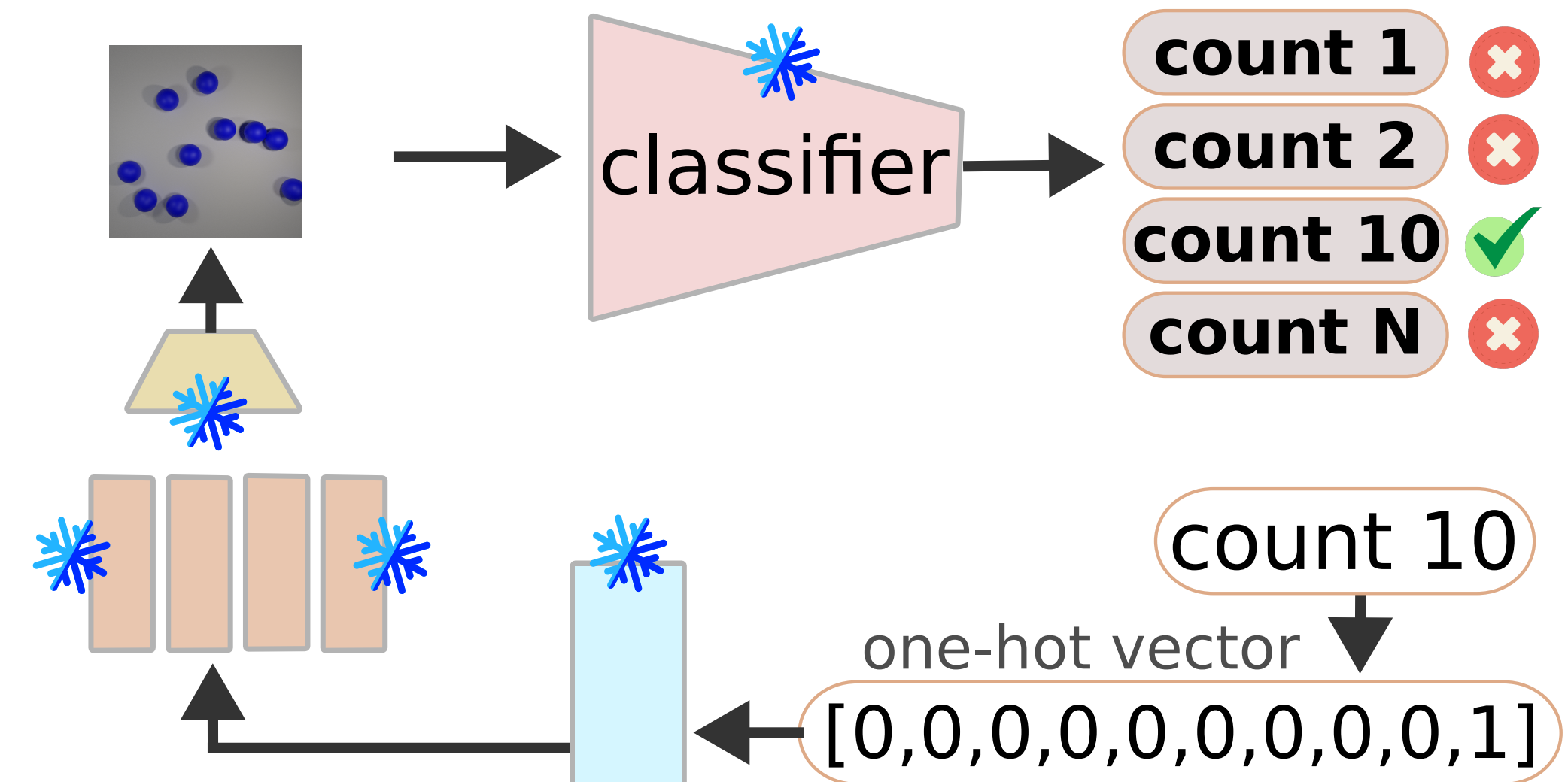
Experimental Settings

Training and Evaluation Pipeline

- **Training stage:** Train conditional diffusion models (U-Net, DiT)



- **Evaluation stage:** Generate images and evaluate generation accuracy with a pretrained classifier



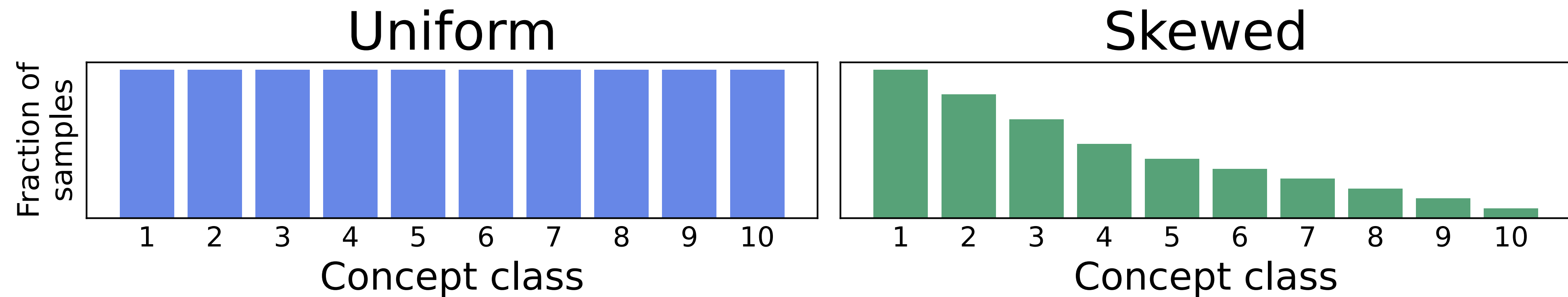
Hypothesis 1

Does concept imbalance affect generation?

Hypothesis 1

Does concept imbalance affect generation?

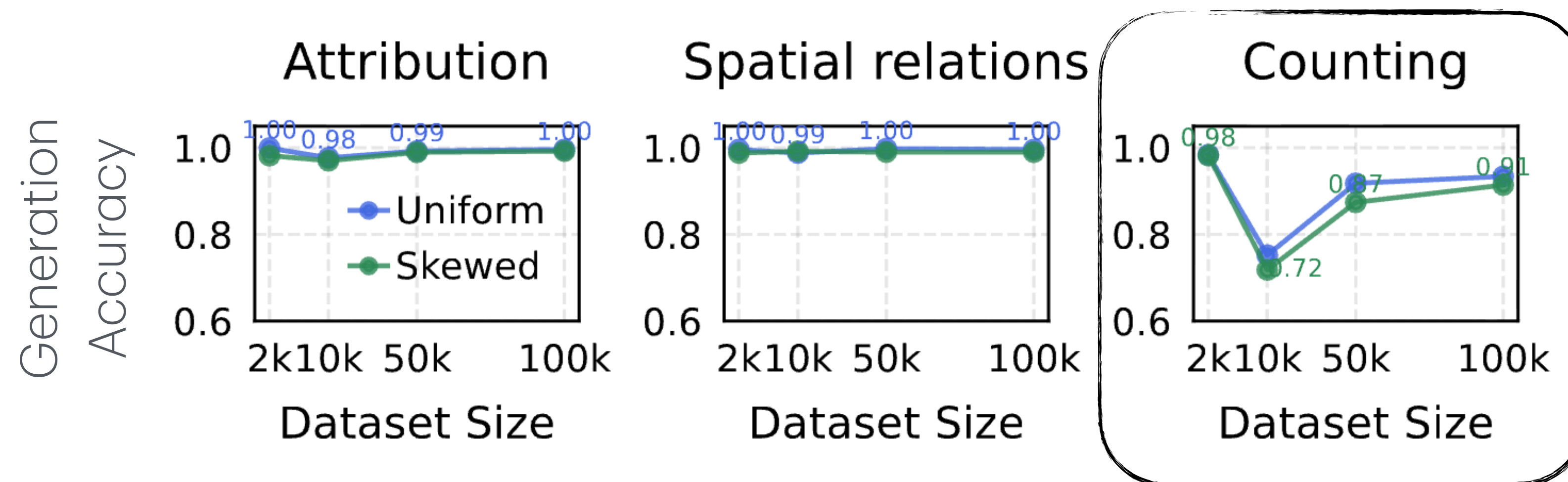
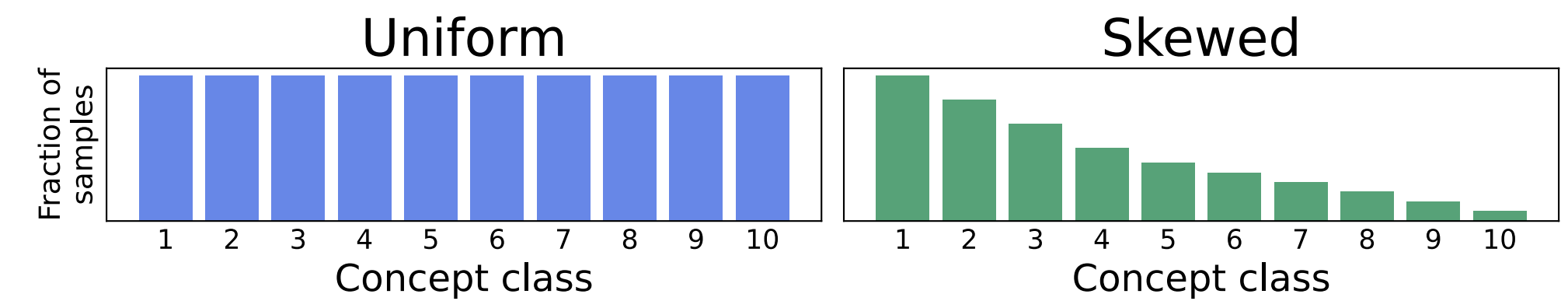
- We vary **total dataset size** and **concept distribution** to test the effect of data imbalance in training.



Hypothesis 1

Does concept imbalance affect generation?

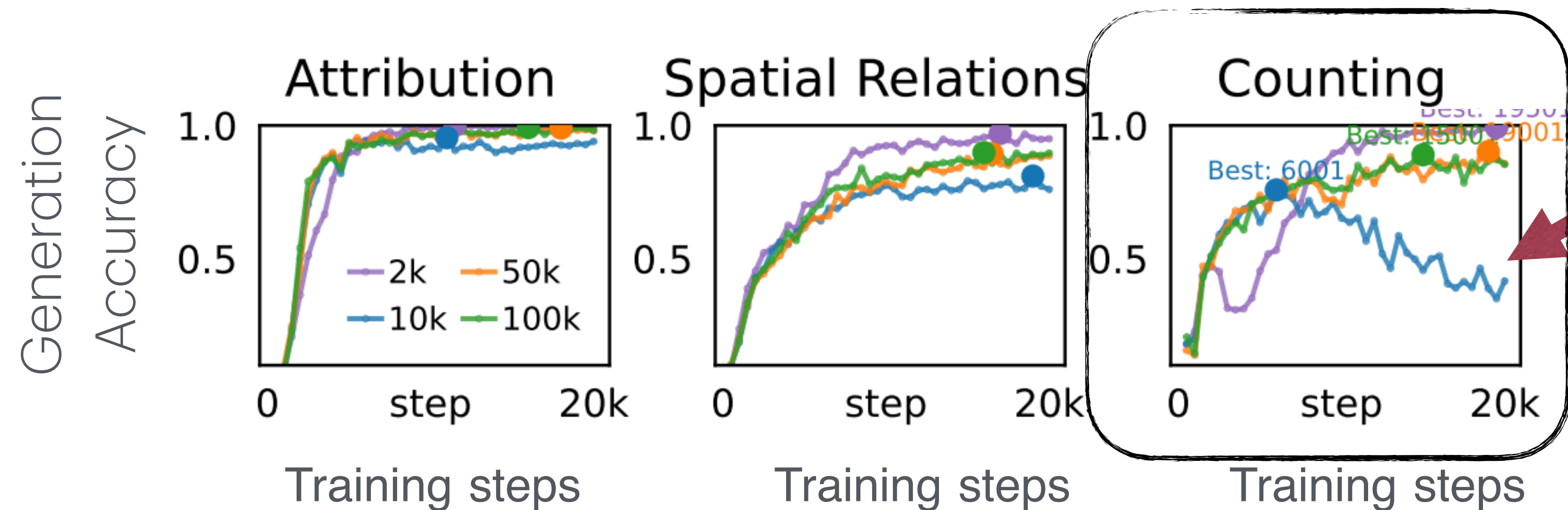
- Data imbalance has limited effect, while **dataset size** matters more.
- Counting is most affected in the low-data regime (e.g., 10k).



Hypothesis 1

Does concept imbalance affect generation?

- Counting is more unstable during training than attribution or spatial relations.



Hypothesis 1

Does concept imbalance affect generation?

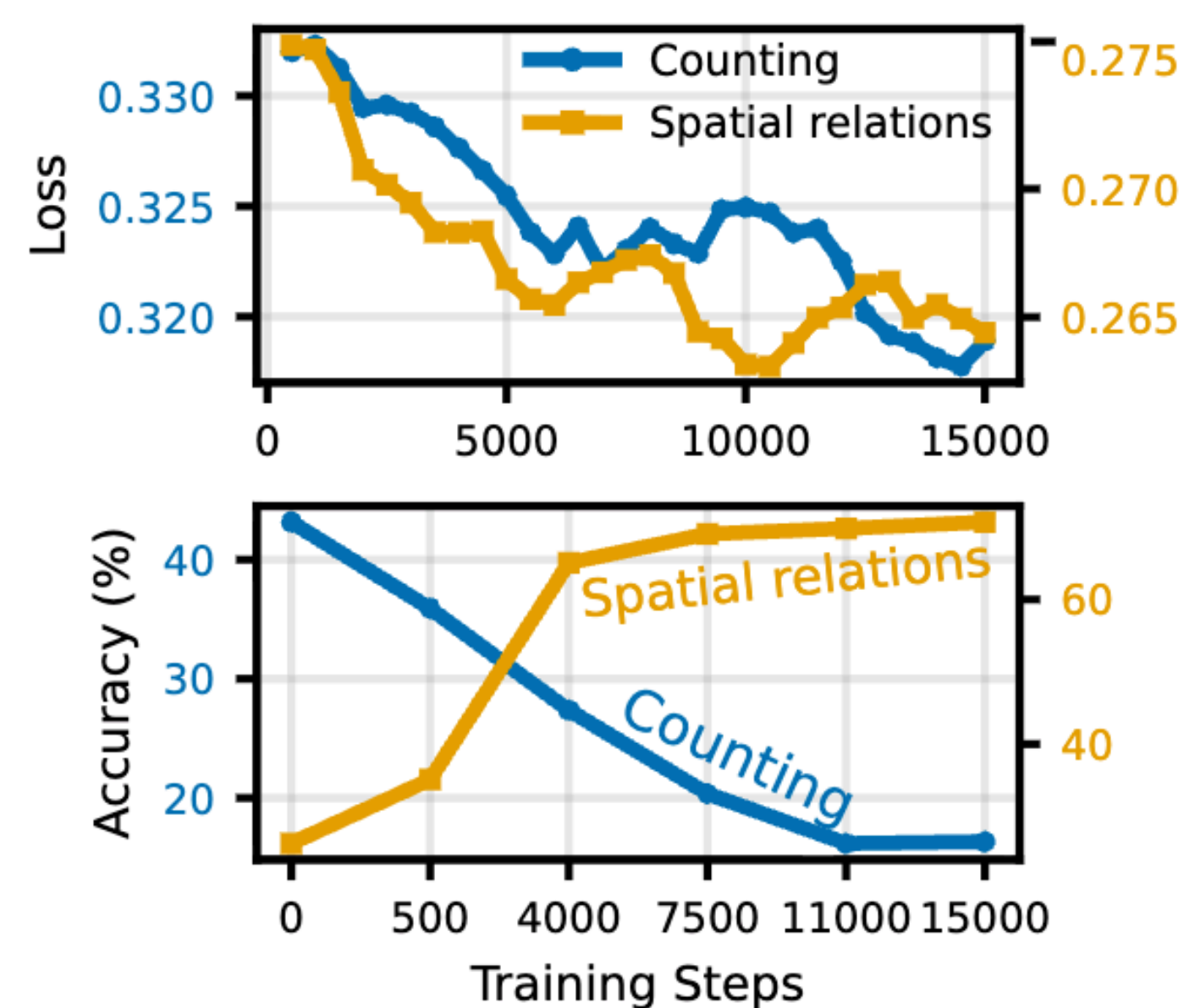
- **Takeaway1:** With sufficient data, models can learn multi-object concepts. In low-data regimes, however, Counting struggles with stable training.

Hypothesis 1

Does concept imbalance affect generation?

- **Takeaway1:** With sufficient data, models can learn multi-object concepts. In low-data regimes, however, Counting struggles with stable training.

SD-3 Finetuning with LoRA



Generated examples



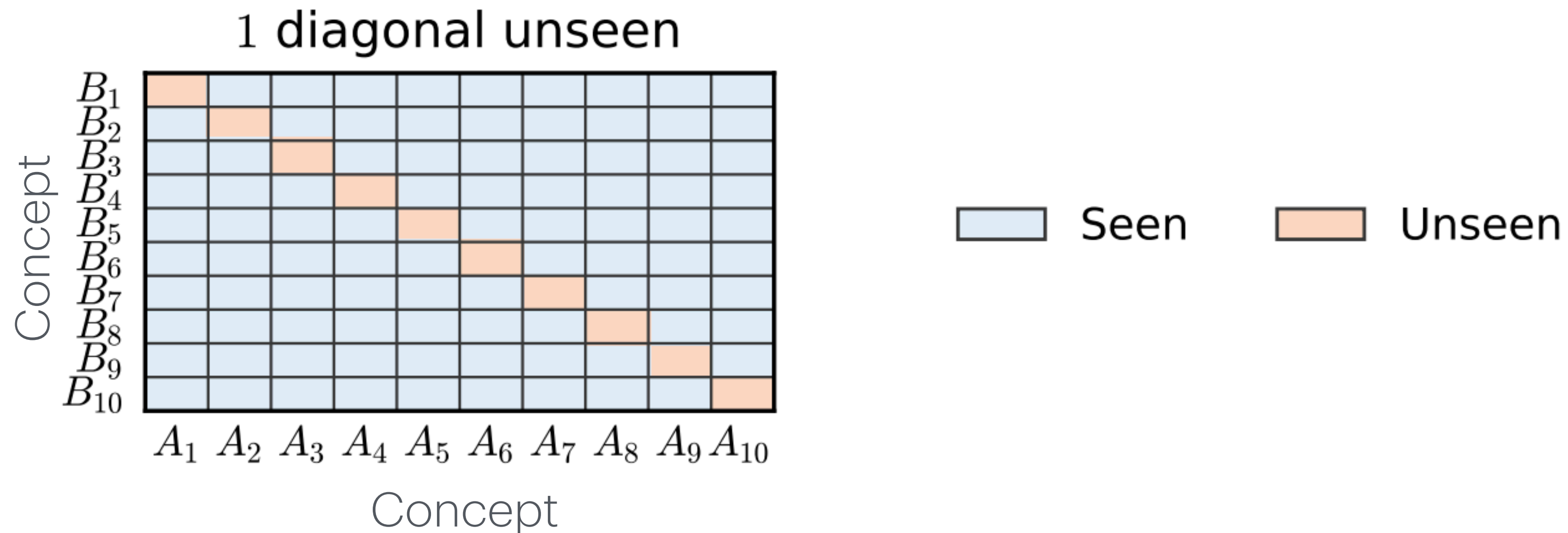
Hypothesis 2

Can models generalize to unseen compositions?

Hypothesis 2

Can models generalize to unseen compositions?

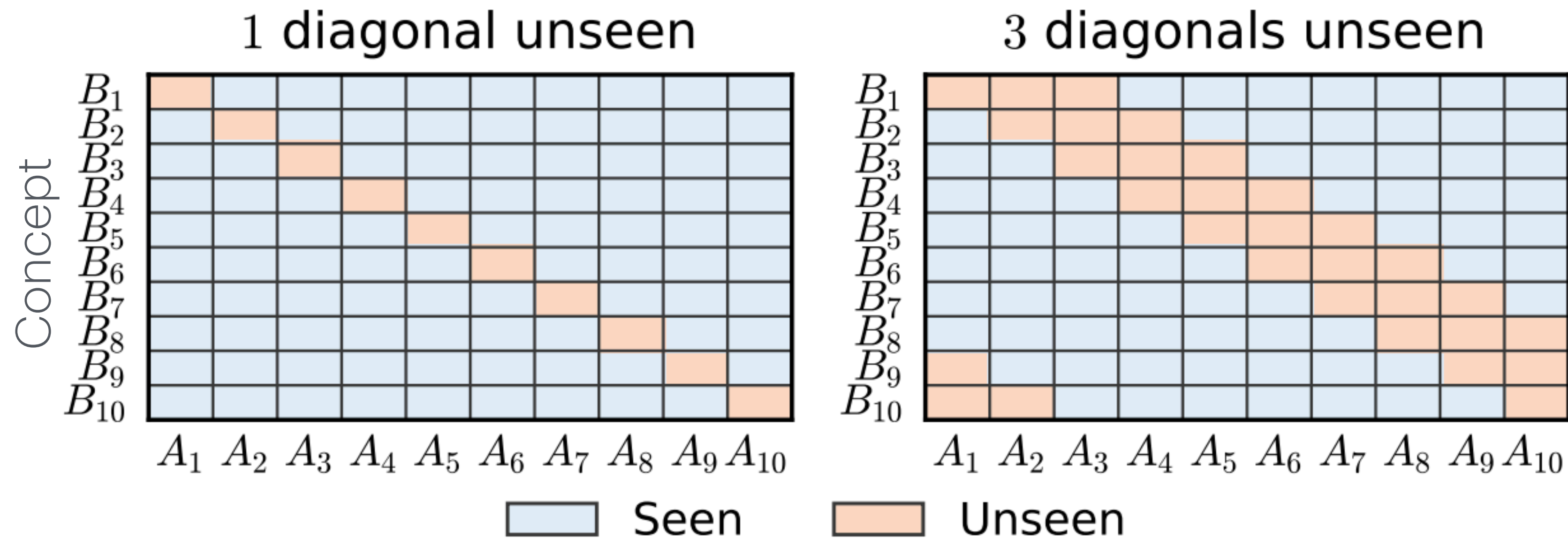
- We vary **total dataset size** and **the number of held-out composition diagonals** in training: 0, 1, 3, 5, and 8.



Hypothesis 2

Can models generalize to unseen compositions?

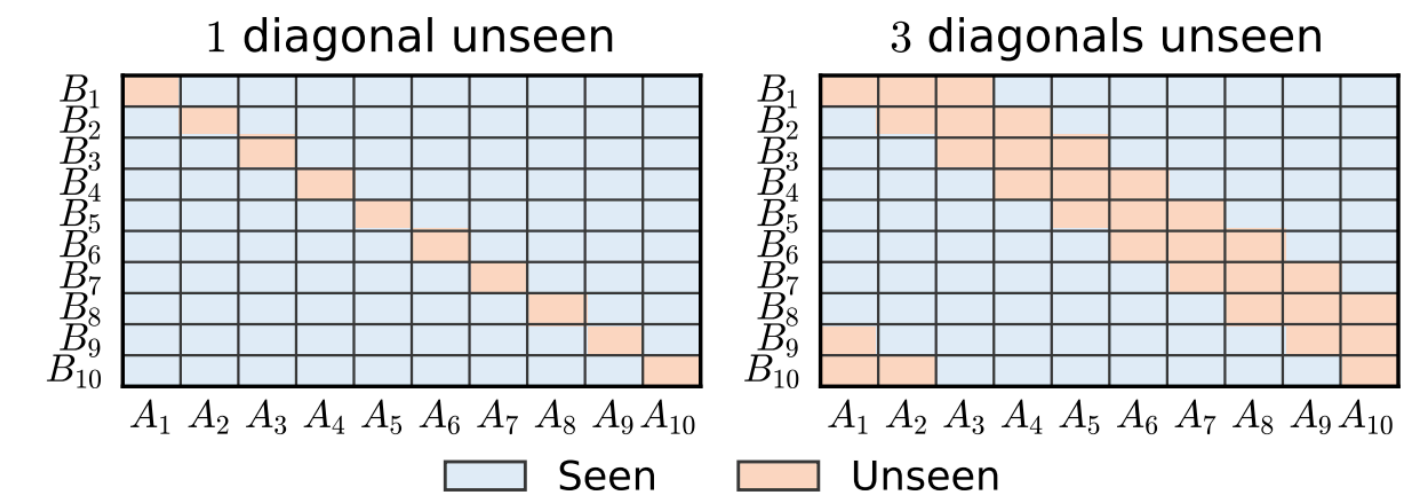
- We vary **the number of held-out composition diagonals** in training: 0, 1, 3, 5, and 8.



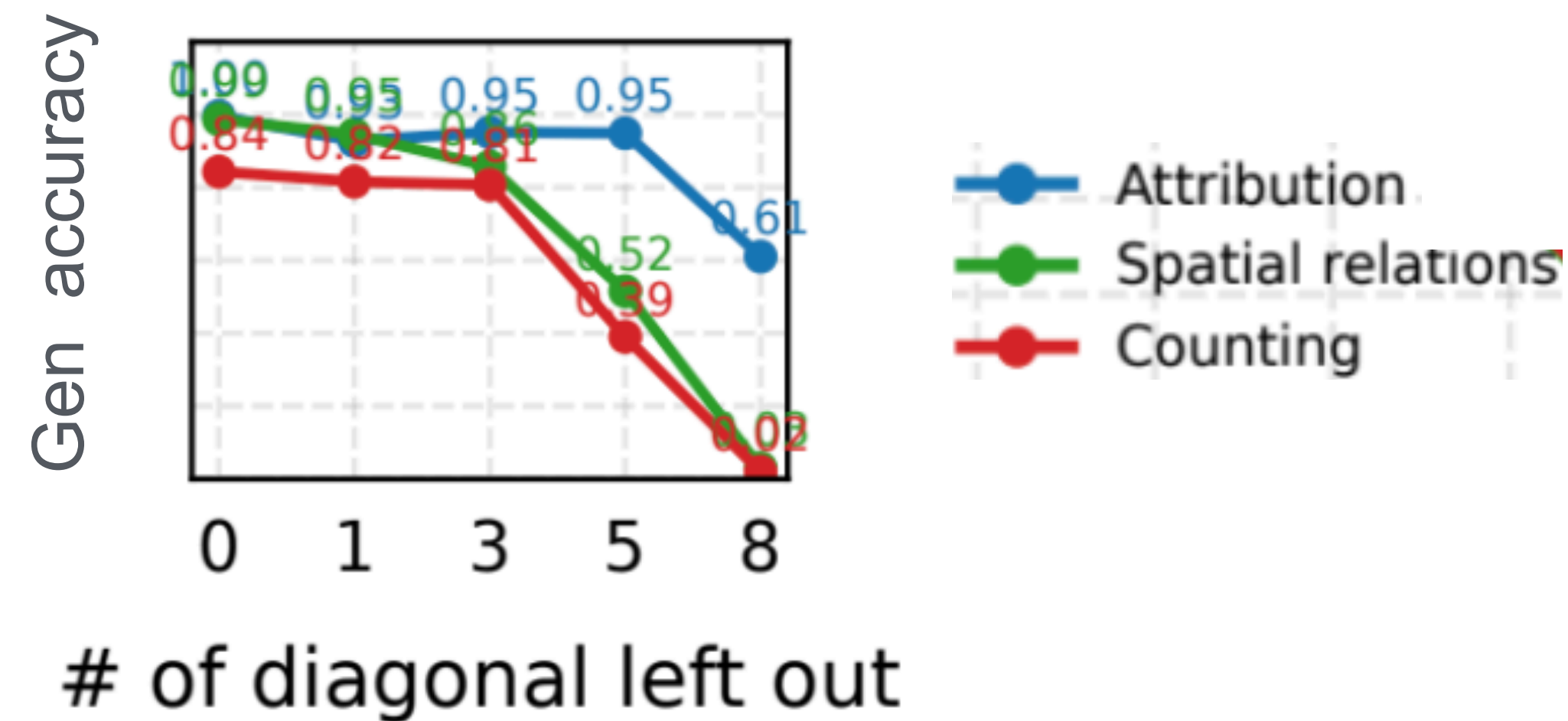
Hypothesis 2

Can models generalize to unseen compositions?

- As more compositions are held out during training, generalization (accuracy) to unseen compositions becomes harder.



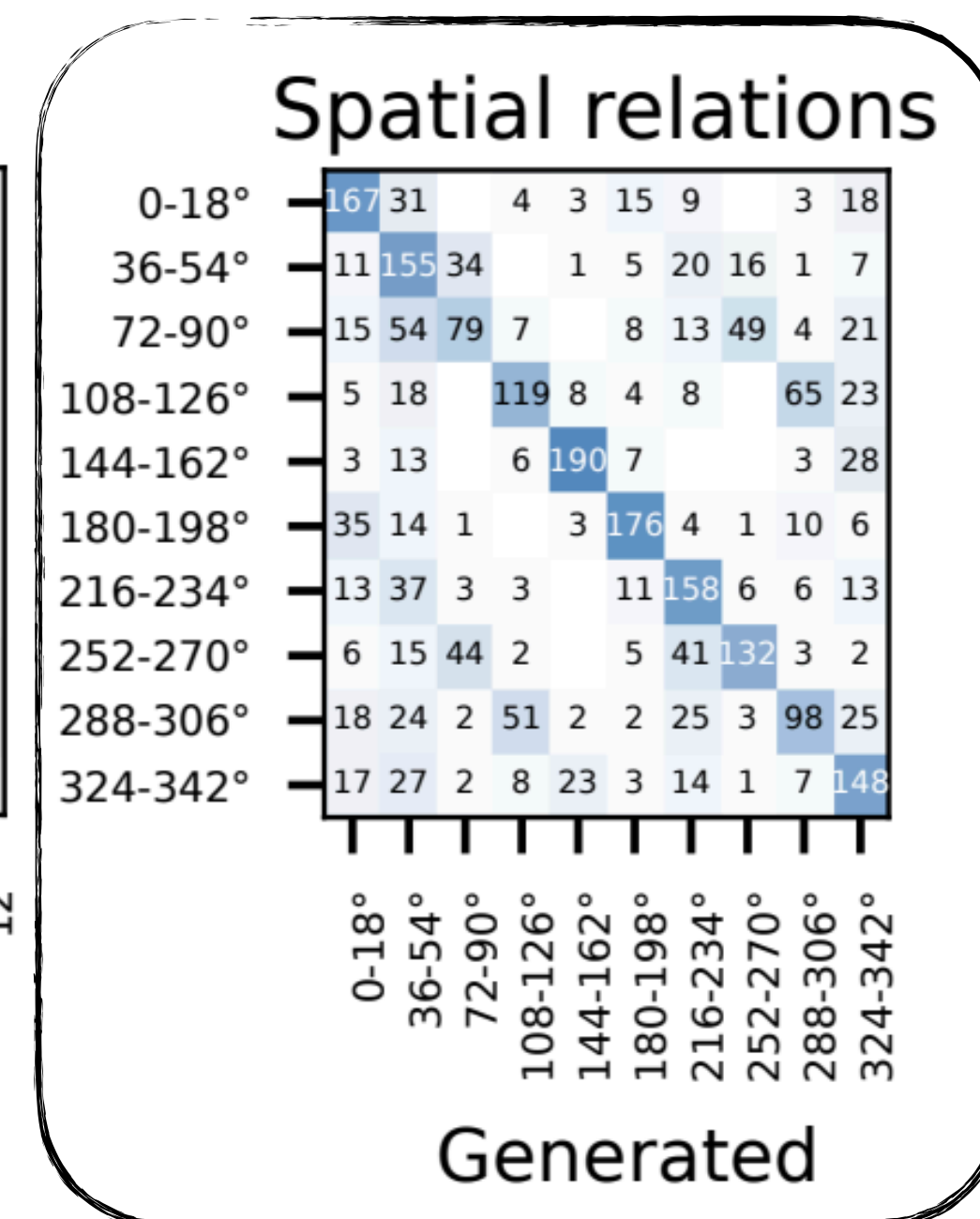
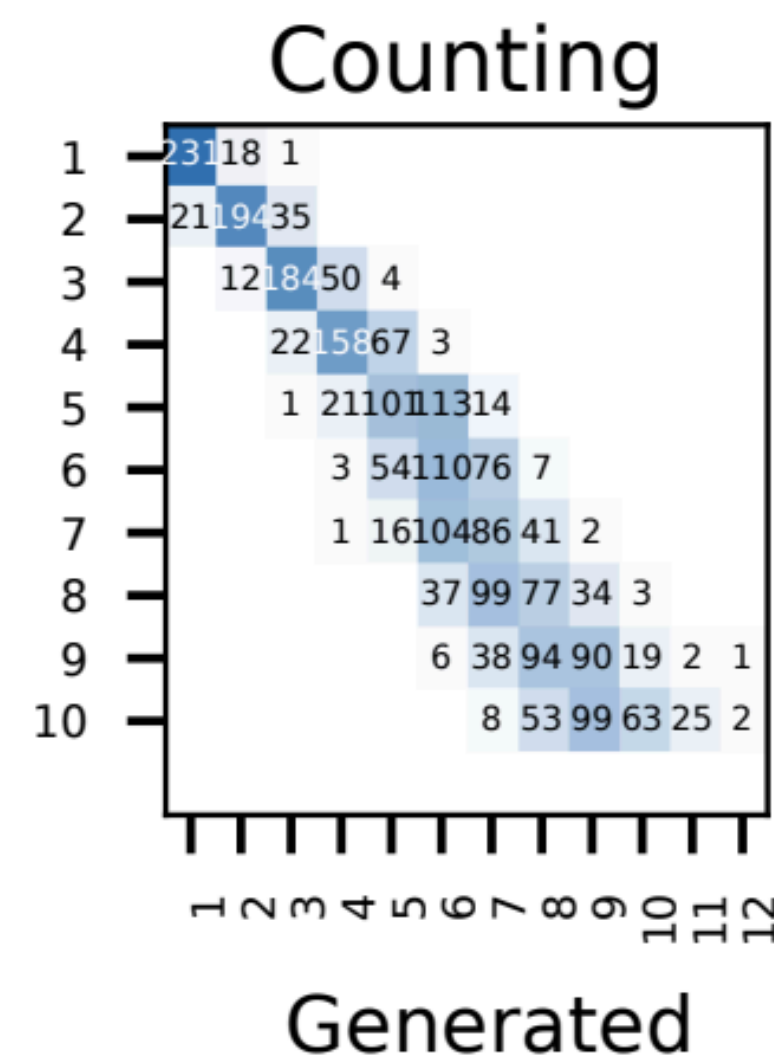
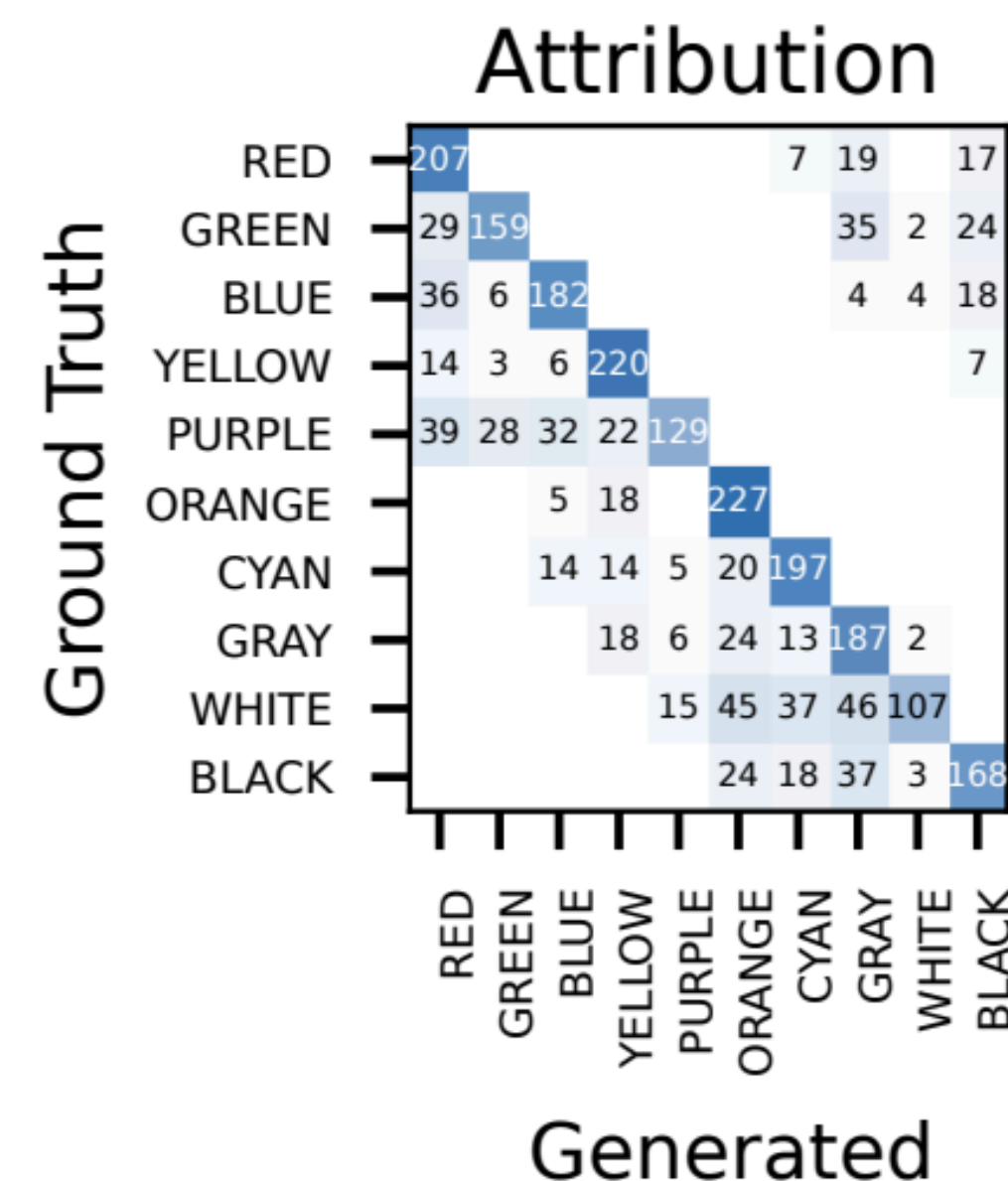
Unseen compositions



Hypothesis 2

Can models generalize to unseen compositions?

- Case study: confusion matrix with five held-out diagonals.
- Spatial relations shows the broadest confusion across generated classes, followed by Counting and Attribution.



Hypothesis 2

Can models generalize to unseen compositions?

- **Takeaway2:** Compositional generalization degrades as more concept combinations are held out during training. The difficulty follows: Attribution < Counting < Spatial Relations.

Thank you!

More results are available in the paper! 😊

Email to : yujin.jeong@tu-darmstadt.de

Code: <https://github.com/eugene6923/MOSAIC>