Imperial College London

Measuring axiomatic soundness of counterfactual image models

Miguel Monteiro, Fabio de Sousa Ribeiro, Nick Pawlowski, Daniel Coelho de Castro, Ben Glocker

BioMedIA

Imperial College London

Fabio de Sousa Ribeiro

Nick Pawlowski

Daniel Coelho de Castro

Ben Glocker

BioMedIA

Summary

- Motivation;
- Background;
- Methods;
- Experiments and results;
- Conclusion.

Motivation

Image Counterfactuals

Motivation

- Counterfactuals can be useful for explainability, interpretability fairness and data-augmentation;
- To generate counterfactuals we must know the data’s generative model aka the **mechanism**;
- When the true mechanism is not known we can estimate an approximation from data;
- In the case of images, the true mechanism is usually not available;
- Additionally, deep generative models are essential to estimate image mechanisms due to their complexity.

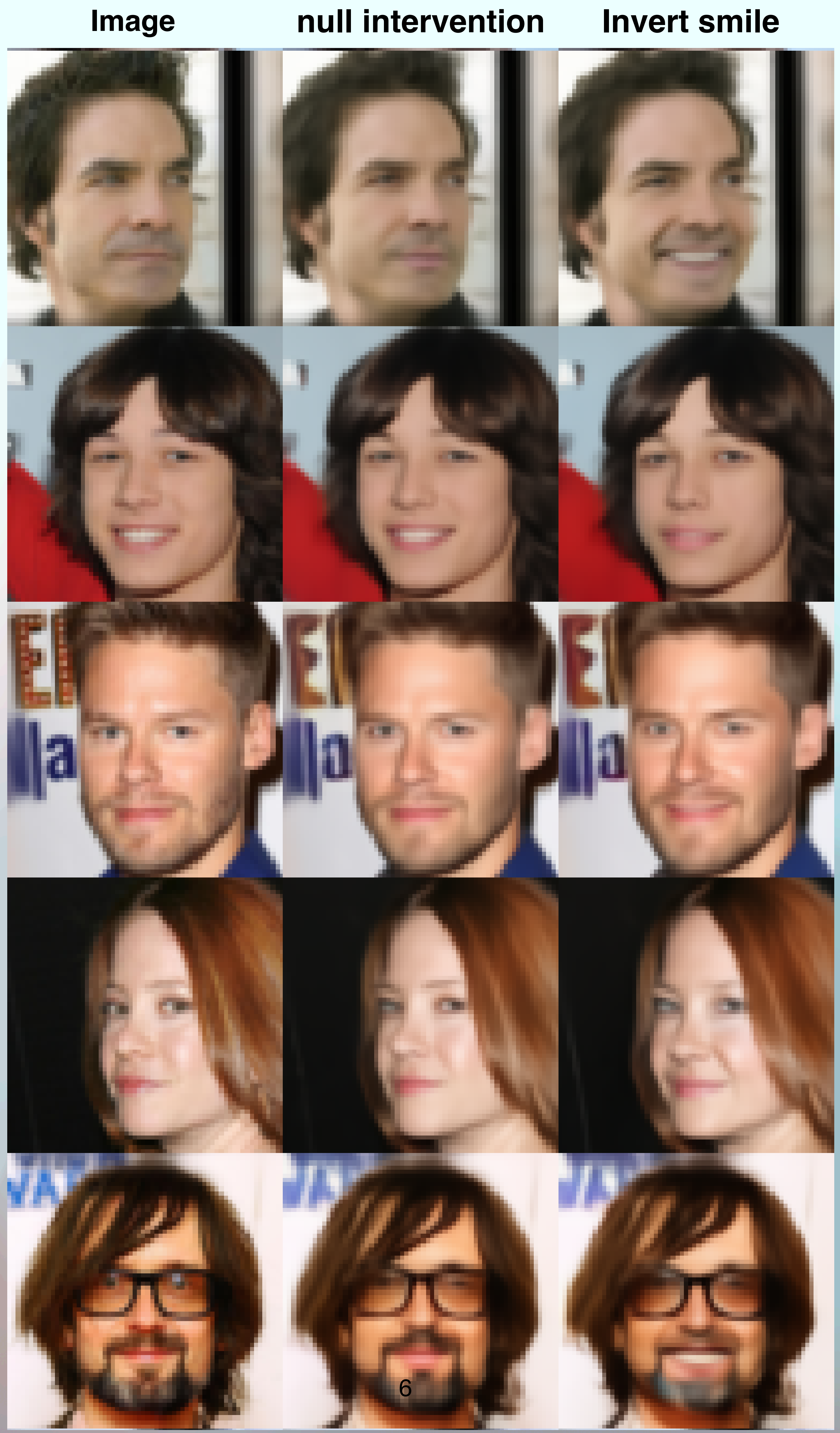


Image
null intervention
Invert smile

Measuring soundness of approximate counterfactuals

Motivation

- We will see that in general deep generative models are only able approximate causal models leading;
- Approximate models lead to approximate estimates for causal effects and counterfactuals;
- Many approaches have been proposed for approximate counterfactual inference in images;
- Less work has been done on evaluating the quality of these approximations;
- This is what our work focuses on.

Background

Counterfactuals (1)

Background

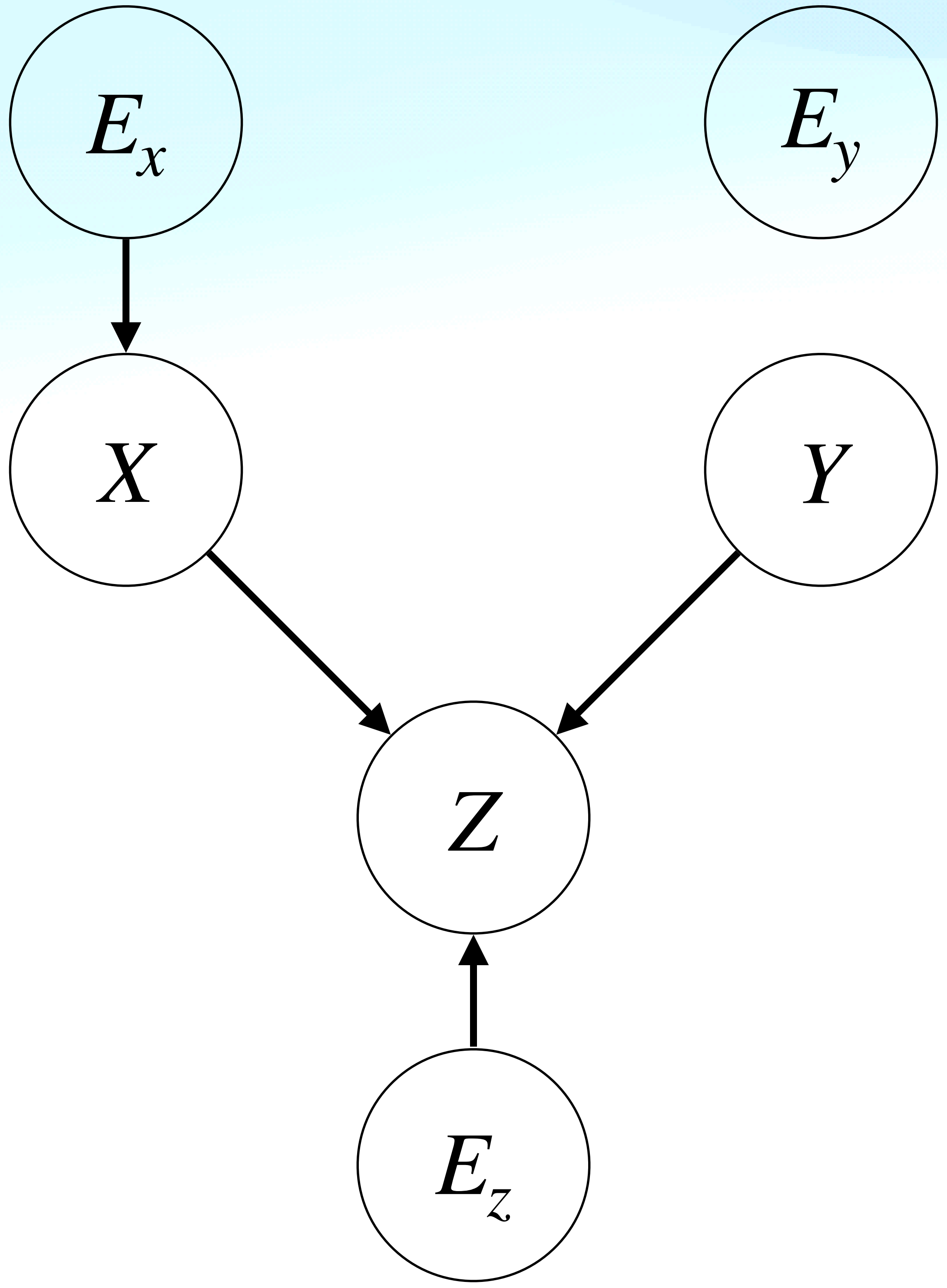
- Consider a model $x = g(\epsilon, \mathbf{pa})$, where $g(\,\cdot\,)$ is a mechanism that generates an observation x from its endogenous causes (parents) $\mathbf{pa}$, and its exogenous causes ϵ.
- A counterfactual is computed in three steps:
  1. Abduction: calculate $p(\epsilon \,|\, x, \mathbf{pa})$ by inverting the mechanism $\epsilon = g^{-1}(x, \mathbf{pa})$;
  2. Action: intervene on the parents $\mathbf{Pa} := \mathbf{pa}*$;
  3. Prediction: propagate the effect through the SCM $x* = g(\epsilon, \mathbf{pa}*)$.

Counterfactuals (2)

Background

$$\epsilon_x, \epsilon_y, \epsilon_z \sim p(\epsilon_x, \epsilon_y, \epsilon_z \,|\, x, y, z)$$

$$p(z \,|\, x, do(y), \epsilon_x, \epsilon_y, \epsilon_z) = \,?$$



Model identifiability

Background

- Assume $P_\theta(X)$ is a distribution of some random variable X, θ is its parameter that takes values in some parameter space Ω_θ . Then, if $P_\theta(X)$ satisfies $p_{\theta_1}(X) \neq p_{\theta_2}(X) \iff \theta_1 \neq \theta_2 \;\forall\; \theta_1, \theta_2 \in \Omega_\theta$, we say that P_θ is identifiable w.r.t. θ on Ω_θ;
- In simple terms, different model parameters must result in different observational distributions.

Model identifiability in deep models (1)

Background

- Deep “causal” generative models are simply deep latent generative models where the latent variables in the model take the role of the exogenous noise;
- The deep mechanism $x = g(\epsilon, \mathbf{pa})$ is coupled with a deep inference model $\epsilon = q(x, \mathbf{pa})$;
- Different model types (e.g. VAEs, GANs, generative flows, diffusion models) will have different choices on how this idea is implemented but the gist is the same.

Model identifiability in deep models (2)

Background

- In the general case, deep models are not identifiable because there are multiple solutions for θ that result in same observational distribution $p_\theta(x \mid \epsilon, \mathbf{pa})$ (Locatello 2020);
- This makes abduction is impossible since $p(\epsilon \mid x, \mathbf{pa})$ is not unique. We can arbitrarily transform ϵ, and, as long as we change the parameters θ, we can recover the same observational distribution;

Model identifiability in deep models (3)
Background

- Even if the model was identifiable, the true model is only guaranteed to be recovered in the limit of infinite data;
- Deep causal models are thus usually only deep approximate causal models;
- We propose measuring the quality of these approximations.

Model identifiability in deep models (3)

Background

- Deep causal models are thus not really causal and are only approximating causal models;
- We shift

Methods

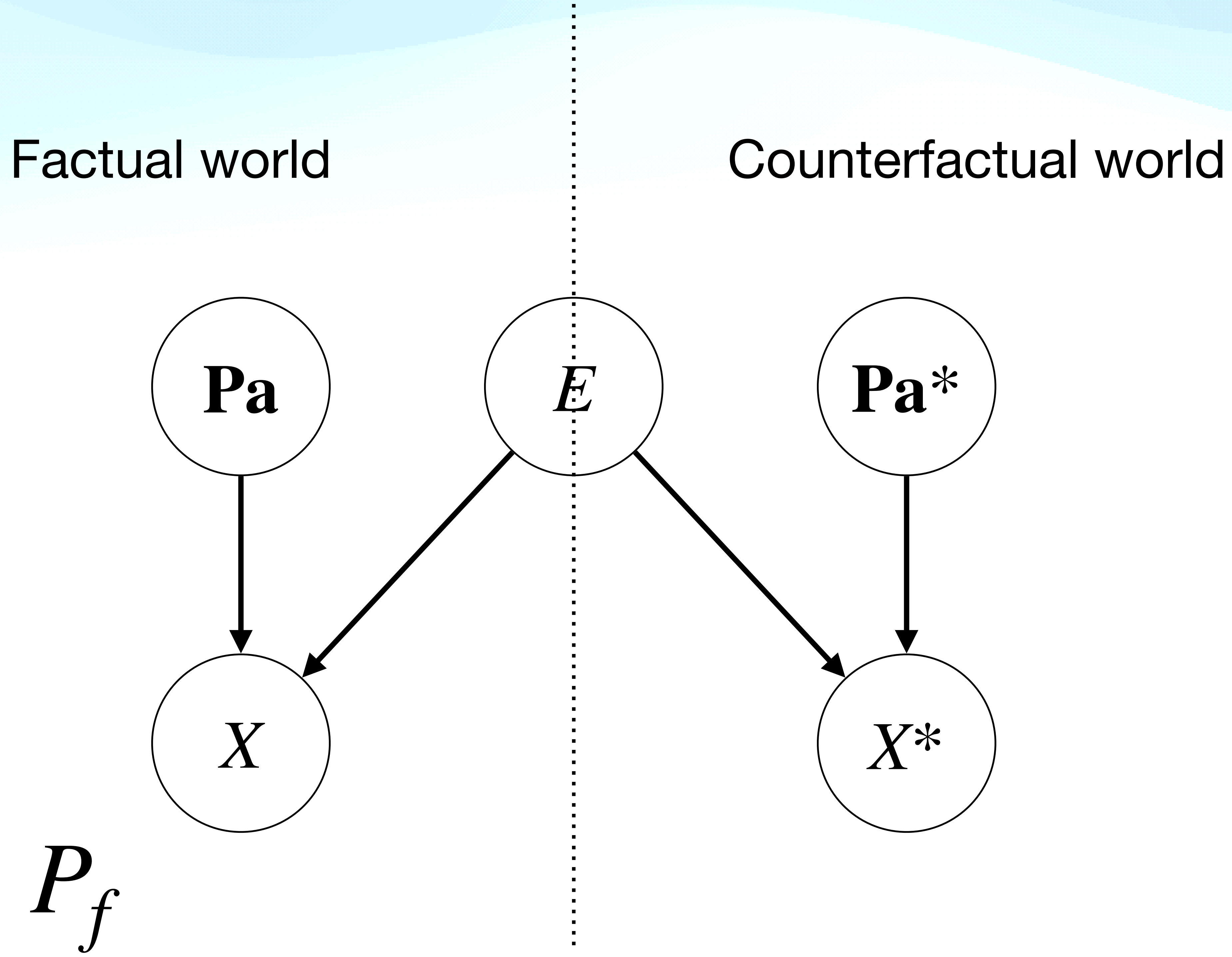
Counterfactual as functions

Methods

- Computationally we can write the three step counterfactual process in one single functional assignment;
- The 3 step process

  1. $\epsilon = \text{abduct}(x, \mathbf{pa})$
  2. $\mathbf{Pa} := \mathbf{pa}*$
  3. $x* = g(\epsilon, \mathbf{pa}*)$

- becomes $x* = f(x, \mathbf{pa}, \mathbf{pa}*)$, where $f \sim P_f$



Counterfactual axioms

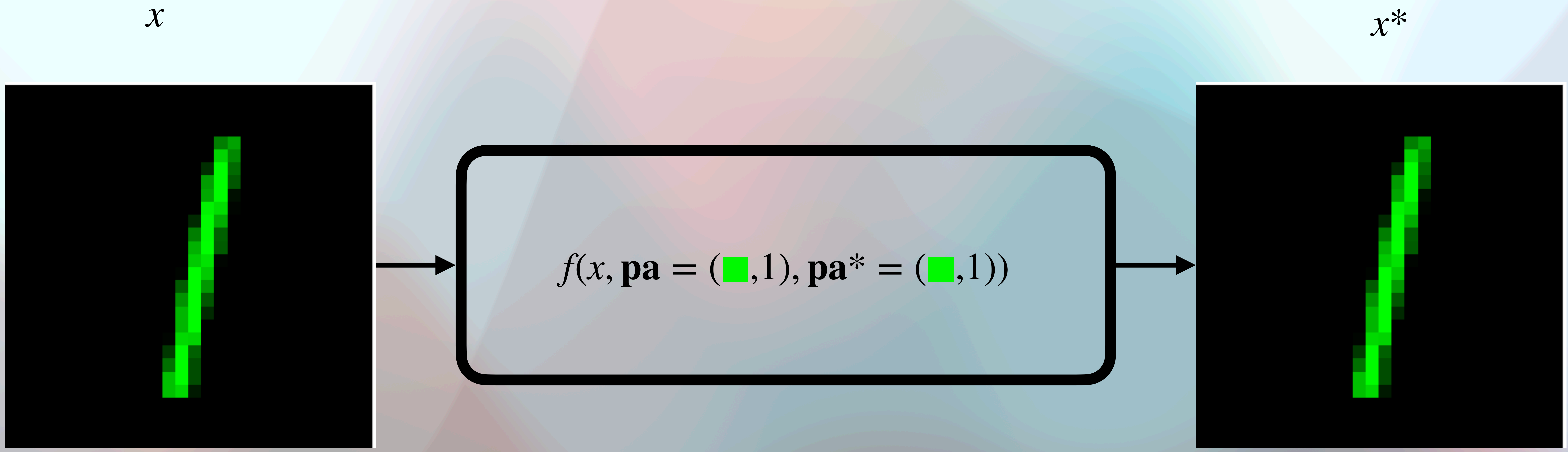
Methods

- The soundness theorem states that the properties of composition, effectiveness and reversibility hold true in all causal models (Galles & Pearl, 1998). The completeness theorem states that these properties are complete (Halpern, 1998);
- Composition, effectiveness and reversibility are the necessary and sufficient properties of counterfactuals in any causal model;
- Evaluating these properties is possible for approximate counterfactuals.

Composition

- Intervening on a variable to have the value it would otherwise have without the intervention will not affect other variables in the system.

- This implies the existence of a null intervention $f(x, \mathbf{pa}, \mathbf{pa}) = x$ since if $\mathbf{pa} = \mathbf{pa}^*$, then x is not affected.



Measuring composition

- To measure composition we can use image distance metrics;
- Given a distance metric $\mathrm{d}_{\mathrm{X}}(\,\cdot\,,\,\cdot\,)$, such as the l_1 distance, an observation x with parents $\mathbf{pa}$ and a functional power m, we can measure composition as:

$$\text{composition}^{(m)}(x, \mathbf{pa}) := \mathrm{d}_{\mathrm{X}}(x, \hat{f}^{(m)}(x, \mathbf{pa}, \mathbf{pa})).$$

- For an ideal model this quantity will always be zero regardless of the number of times we apply f.

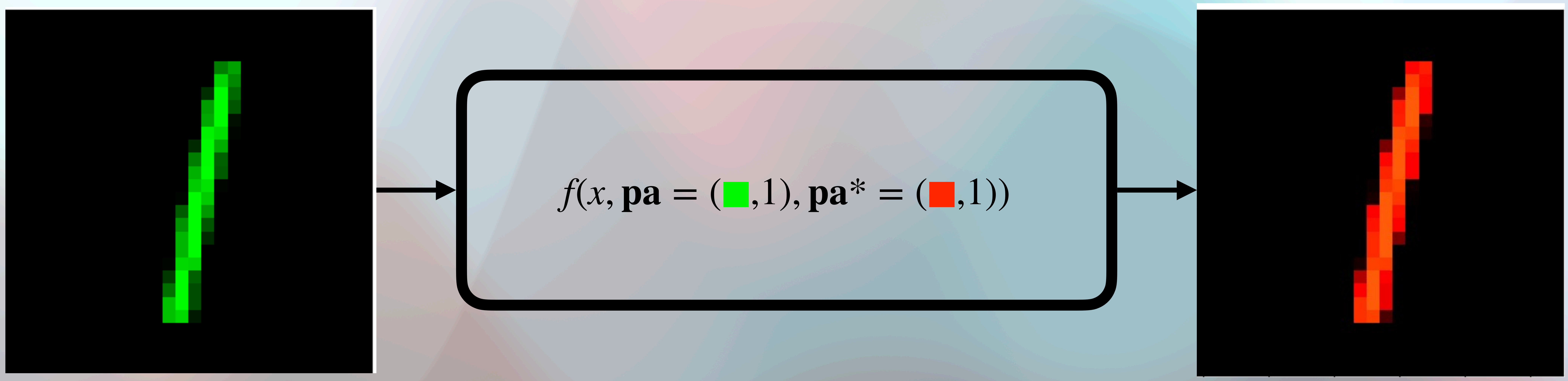
Effectiveness

- Intervening on a variable to have a specific value will cause the variable to take that value.

- Suppose $\mathrm{Pa}(\,\cdot\,)$ is an oracle function that returns the parents of a variable, then we have the following equality:

$$\mathrm{Pa}(f(x, \mathbf{pa}, \mathbf{pa}^*)) = \mathbf{pa}^*$$

x $\qquad$ x^*



$\mathrm{Pa}(x^*) = (red, 1)$

Measuring effectiveness (1)

- Unlike composition, measuring effectiveness is not easy;
- We would like to have an oracle function $\mathrm{Pa}_k(\cdot)$ which returns the value of the parent pa_k given the observation;
- In the absence of this function we approximate it using regressors or classifiers trained from data;
- We must beware that this approximate oracle function is susceptible to confounding of effects and take appropriate measures.

$\mathrm{Pa}(x^*) = (red, 1)$

Measuring effectiveness (2)

- We measure effectiveness individually for each parent by creating a pseudo-oracle function $\widehat{\mathrm{Pa}_k}(\,\cdot\,)$, which returns the value of the parent pa_k given the observation;
- To independently measure how well the effect of each parent is modelled, we measure effectiveness after applying partial counterfactual functions: $\mathbf{pa}^* = \mathbf{pa}_{\mathscr{K}\backslash k} \cup \{\mathrm{pa}_k\}$;
- Using an appropriate distance metric $\mathrm{d_k}(\,\cdot\,,\,\cdot\,)$, we measure effectiveness for each parent as:

$$\mathrm{effectiveness}_k(x, \mathbf{pa}) := \mathrm{d_k}(\widehat{\mathrm{Pa_k}}(\hat{f}(x, \mathbf{pa}, \mathbf{pa}^*)), \mathrm{pa_k}).$$

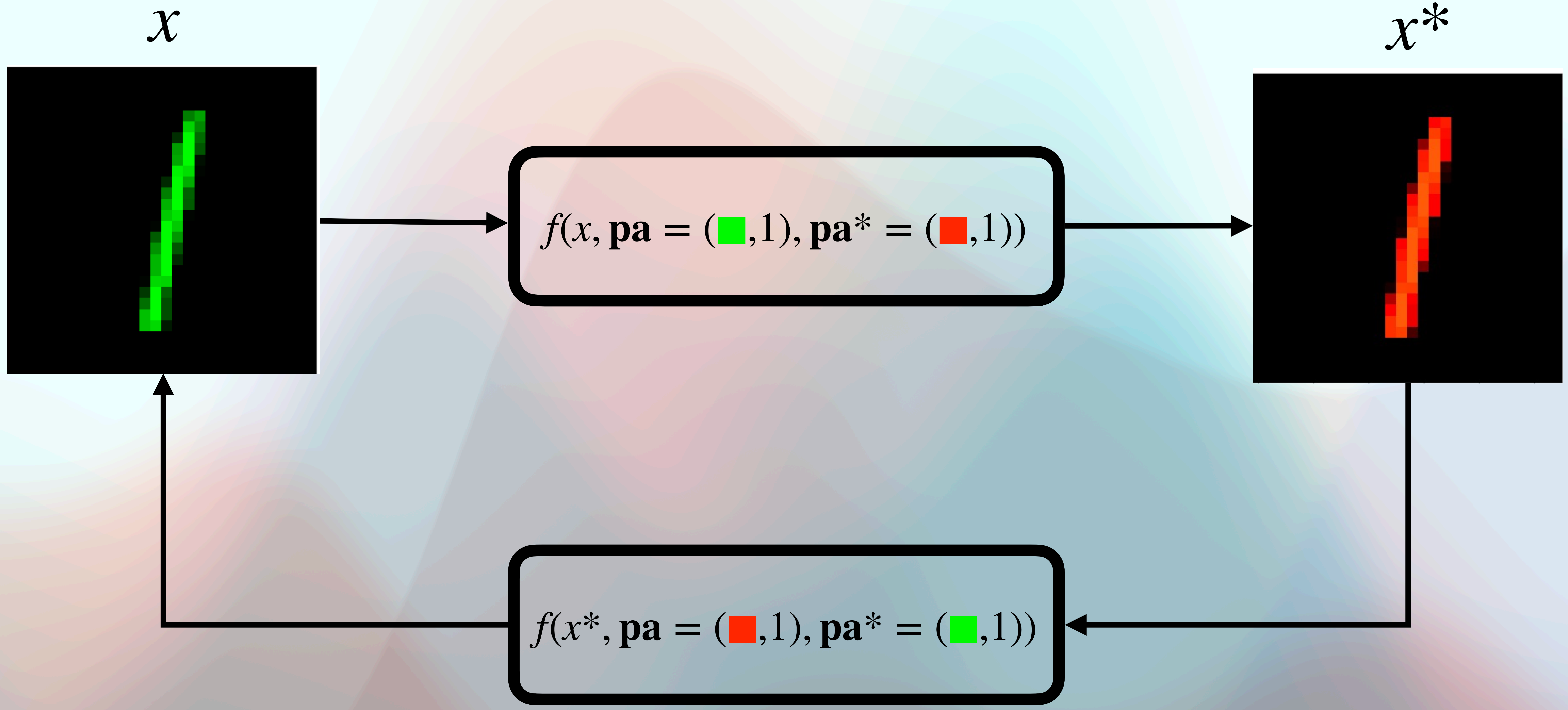
$\mathrm{Pa}(x^*) = (red, 1)$

Reversibility (1)

- If setting a variable X to a value x results in a value y for a variable Y, and setting Y to a value y results in a value x for X, then X and Y will take the values x and y.
- In other words, reversibility prevents the existence of feedback loops;
- In Markovian SCMs, reversibility follows trivially from composition.

Reversibility (2)

- The mapping between the observation and the counterfactual is deterministic for invertible mechanisms.



Measuring reversibility

- Like with composition, we can measure reversibility using image distance metrics (for invertible mechanisms);
- Setting $\hat{p}(x, \mathbf{pa}, \mathbf{pa}*) = \hat{f}(\hat{f}(x, \mathbf{pa}, \mathbf{pa}*), \mathbf{pa}*, \mathbf{pa})$, given a distance metric $\mathrm{d}_\mathrm{X}(\,\cdot\,,\cdot\,)$, such as the l_1 distance, an observation x with parents $\mathbf{pa}$ and a functional power m, we can measure reversibility as

$$\text{reversibility}^{(m)}(x, \mathbf{pa}, \mathbf{pa}*) := \mathrm{d}_\mathrm{X}(x, \hat{p}^{(m)}(x, \mathbf{pa}, \mathbf{pa}*)).$$

- For an ideal model this quantity will always be zero regardless of the number of times we apply p.

Why measure soundness
Methods

- Based on these properties we can measure how far our approximate model is from being truly causal;
- In relation to deep models, we can:
    - compare models without explicit likelihood (GANs);
    - compare models whose performance is disconnected from likelihood, since deep latent variable models can assign arbitrarily high likelihoods to OOD samples (Nalisnick2018).

Experiments and results

Deep generative models as approximate counterfactual functions

Experiments and results

- Any conditional deep latent generative model can be framed as an approximate counterfactual function of the form $x* = \hat{f}_\theta(x, \mathbf{pa}, \mathbf{pa}*)$;
- In this work we look at conditionals variational auto-encoders (VAEs) and generative adversarial networks (GANs).

Conditional VAE

Experiments and results

$$\mathrm{ELBO}_\beta = \mathbb{E}_{q(z|x,\mathbf{pa})}(\log p_\omega(x\,|\,z, \mathbf{pa})) - \beta D_{\mathrm{KL}}(q(z\,|\,x, \mathbf{pa})\,||\,p(z)),$$

where $q_\theta(z\,|\,x, \mathbf{pa})$ is the approximate latent posterior distribution parameterised by a neural network encoder, $p_\omega(x\,|\,z, \mathbf{pa})$ is the conditional observational posterior distribution parameterised by a neural network decoder, and $p(z)$ is the latent prior.

Counterfactuals: 1. $z \sim q_\theta(z\,|\,x, \mathbf{pa})$ 2. $\mathbf{Pa} := \mathbf{pa}*$ 3. $x* \sim p_\omega(x*\,|\,z, \mathbf{pa}*)$

Or rewrite as $x* = \hat{f}_{\theta,\omega}(x, \mathbf{pa}, \mathbf{pa}*)$ where $\hat{f} \sim P_z(\hat{f})$

Conditional GAN with composition constraint

Experiments and results

$$F(\theta, \omega) = \mathbb{E}_{x,\mathbf{pa} \sim P^{do}(x,\mathbf{pa})}[\log D_\omega(x, \mathbf{pa})]$$

$$-\mathbb{E}_{\substack{x,\, \mathbf{pa} \,\sim\, P^{src}(x,\, \mathbf{pa}) \\ \mathrm{pa}_k \,\sim\, P(\mathrm{pa}_k)}}[\log(1 - D_\omega(\hat{f}_\theta(x, \mathbf{pa}, \mathbf{pa}*), \mathbf{pa}))]$$

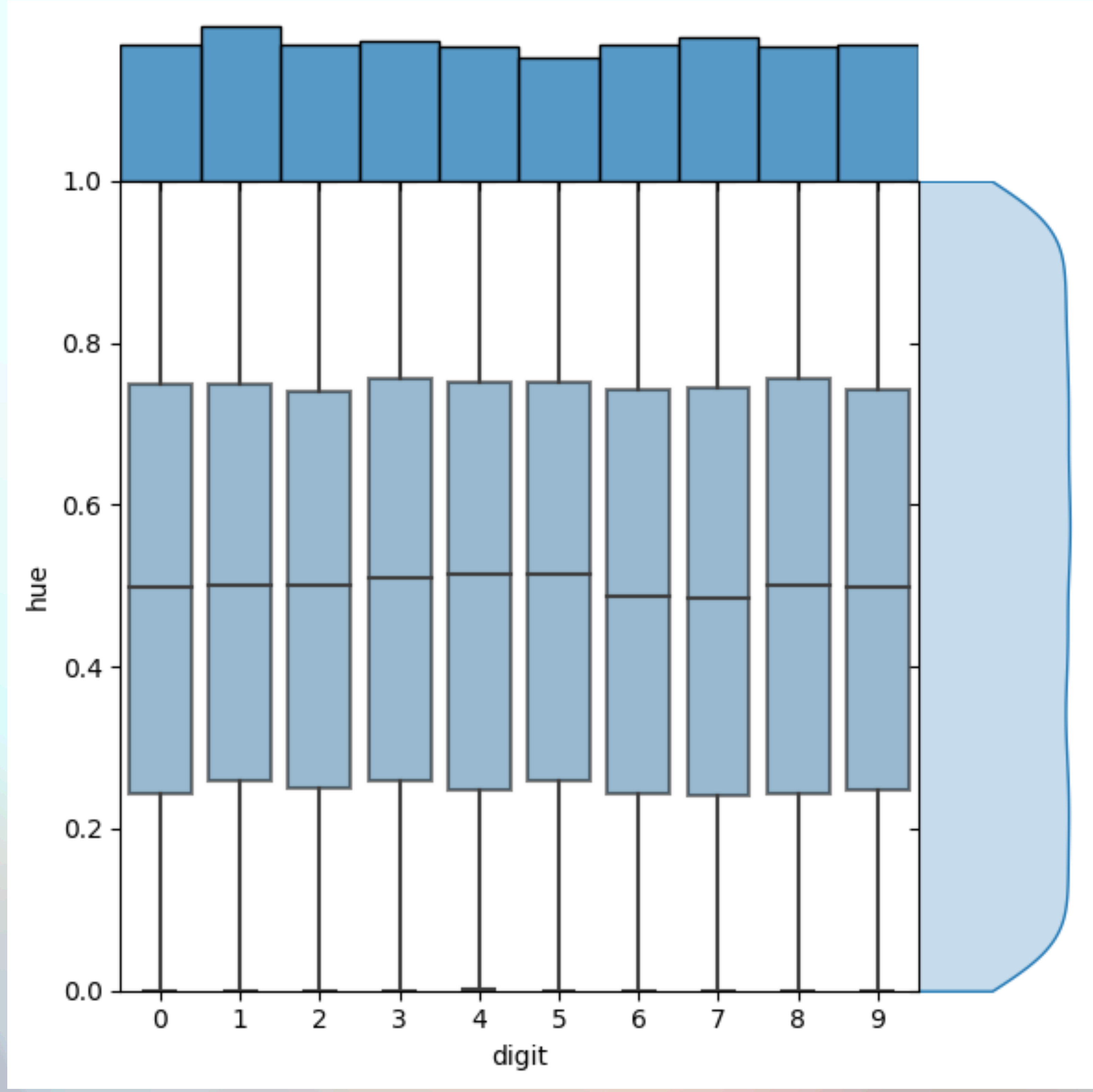
$$+\mathbb{E}_{x,\mathbf{pa} \sim P^{src}(x,\mathbf{pa})} \mathrm{d_X}(x, \hat{f}(x, \mathbf{pa}, \mathbf{pa}))$$

Where the conditional generator $\hat{f}_\theta(x, \mathbf{pa}, \mathbf{pa}*)$ is a neural network approximating the counterfactual function directly.

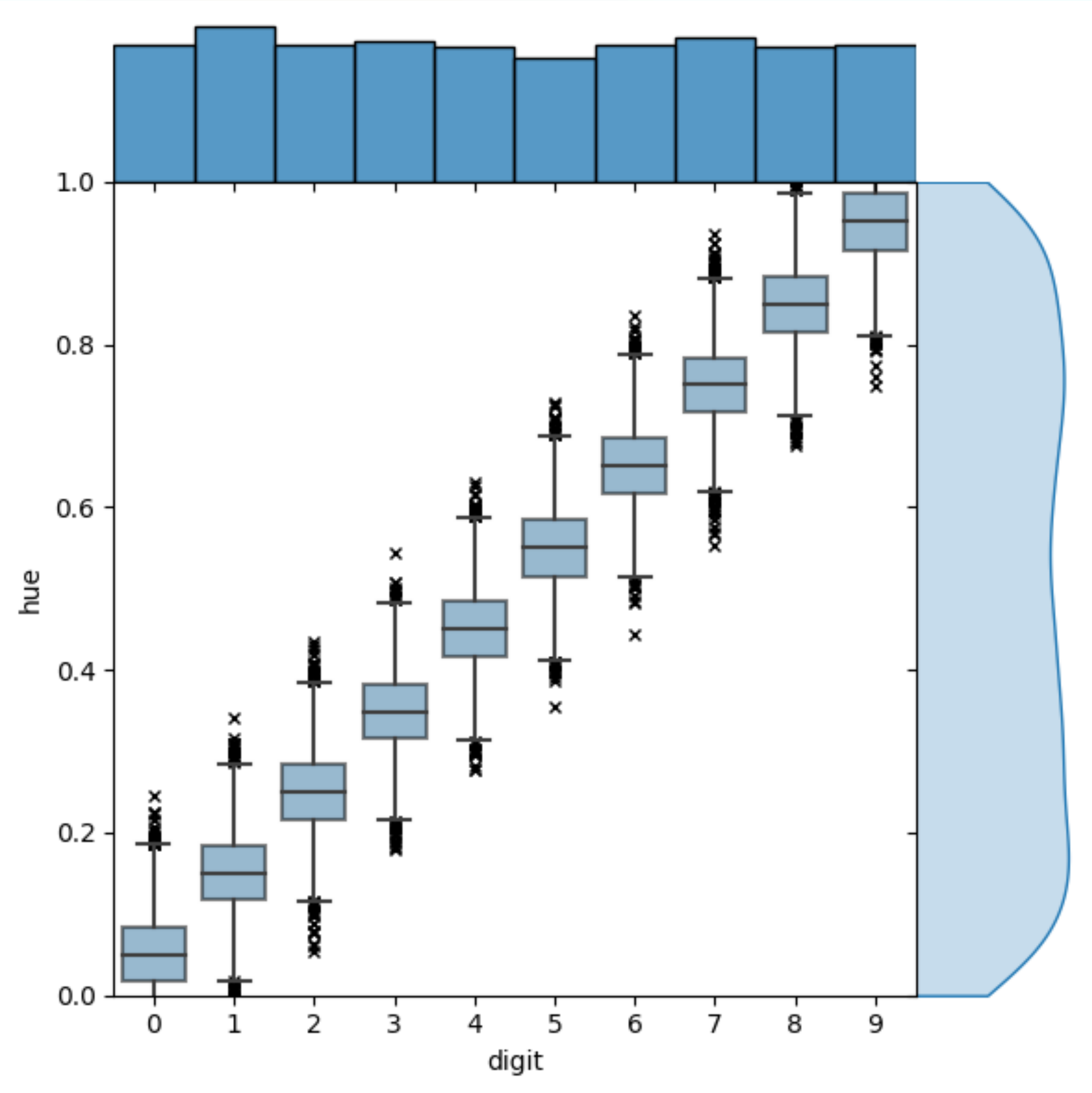
We introduce an additional constraint on the generator to preserve identity (composition).

Colour MNIST (1)

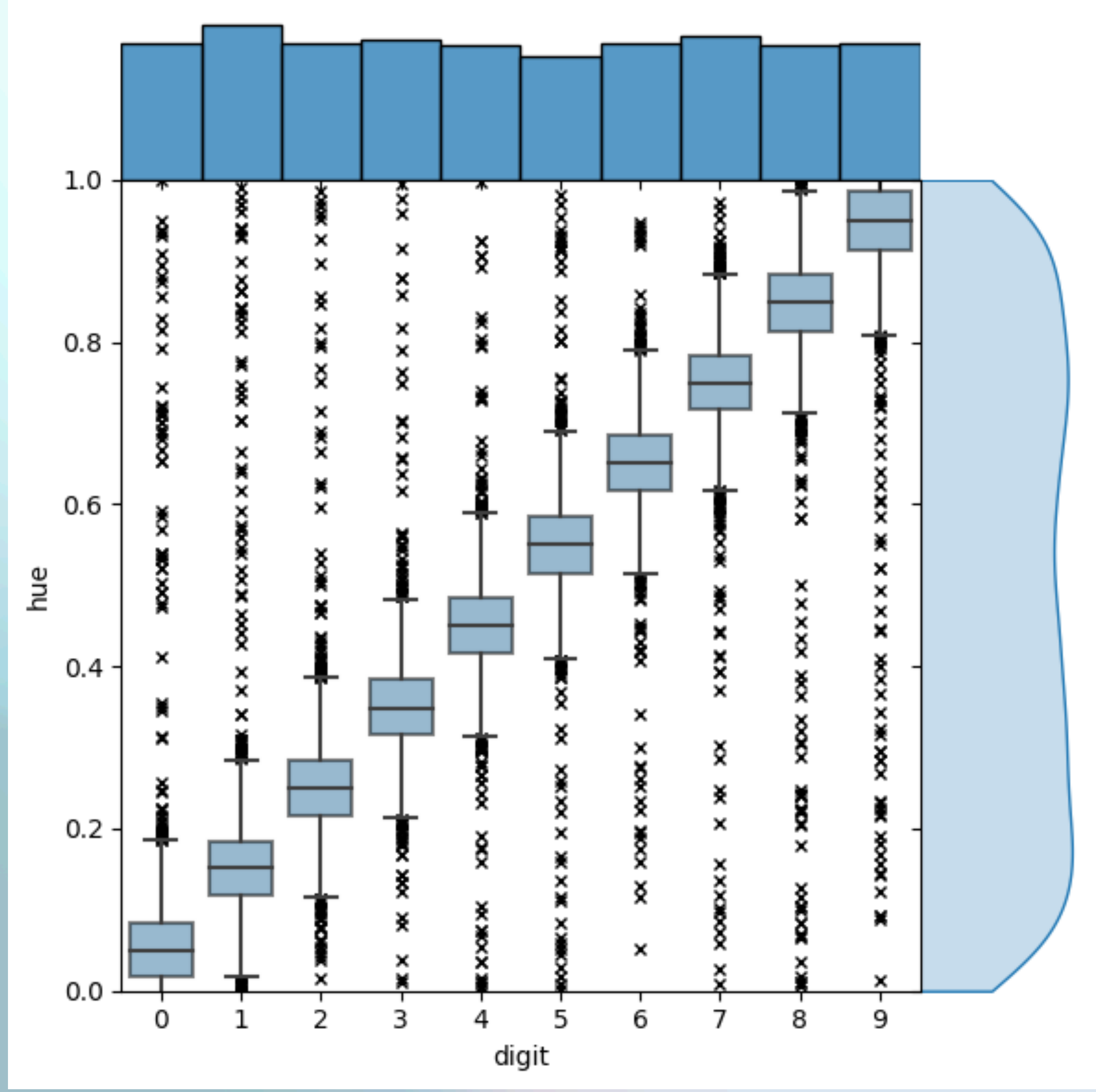
Colour MNIST (2)



Unconfounded joint distribution.



Confounded joint distribution w\o full support.



Confounded joint distribution with full support.

Colour MNIST (3)

Experiments and results

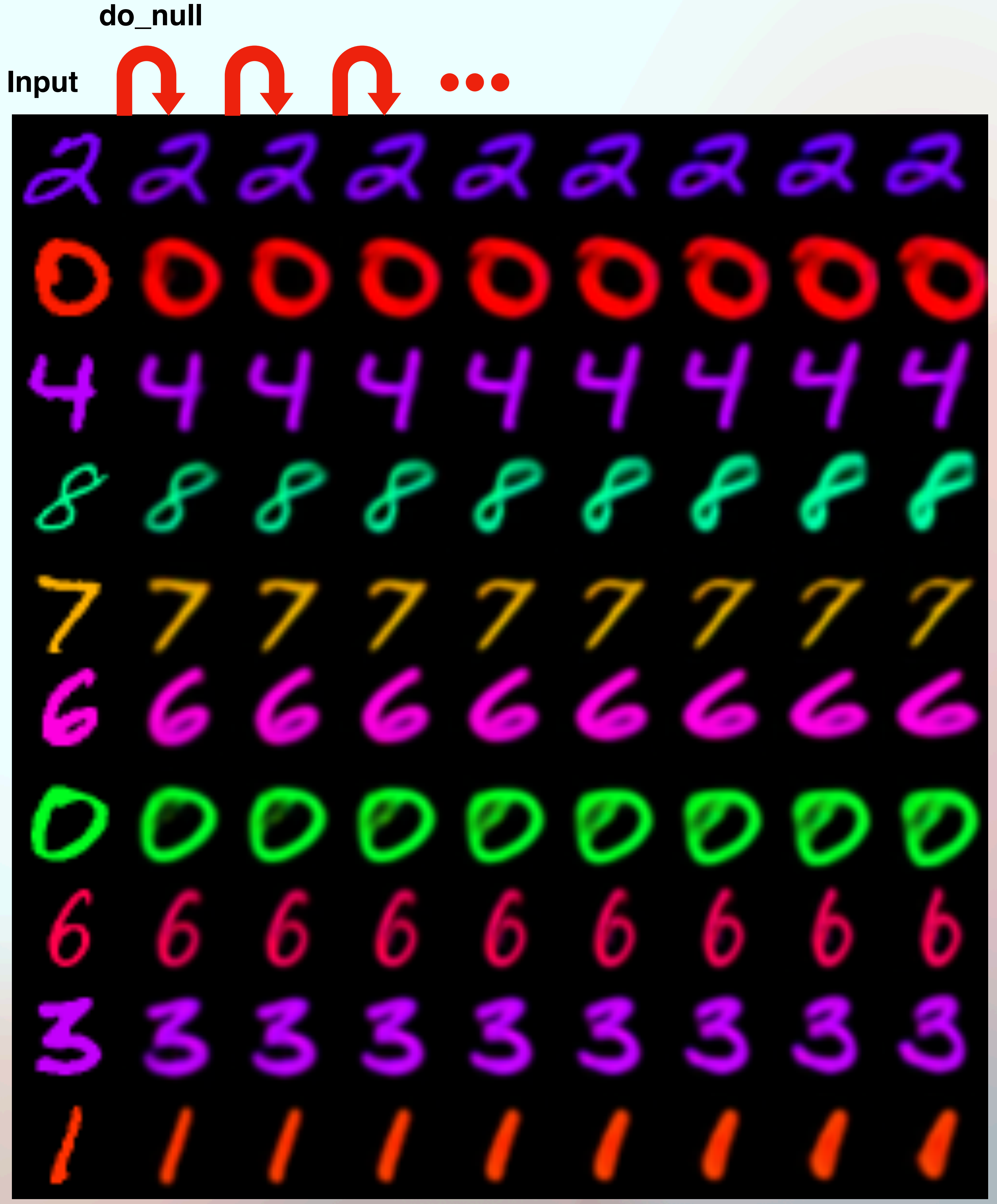
- The goal of the experiment is to see how we can use the derived soundness metrics to compare different models and scenarios visually as well as numerically;

- For demonstration purposes we compare two extreme cases:

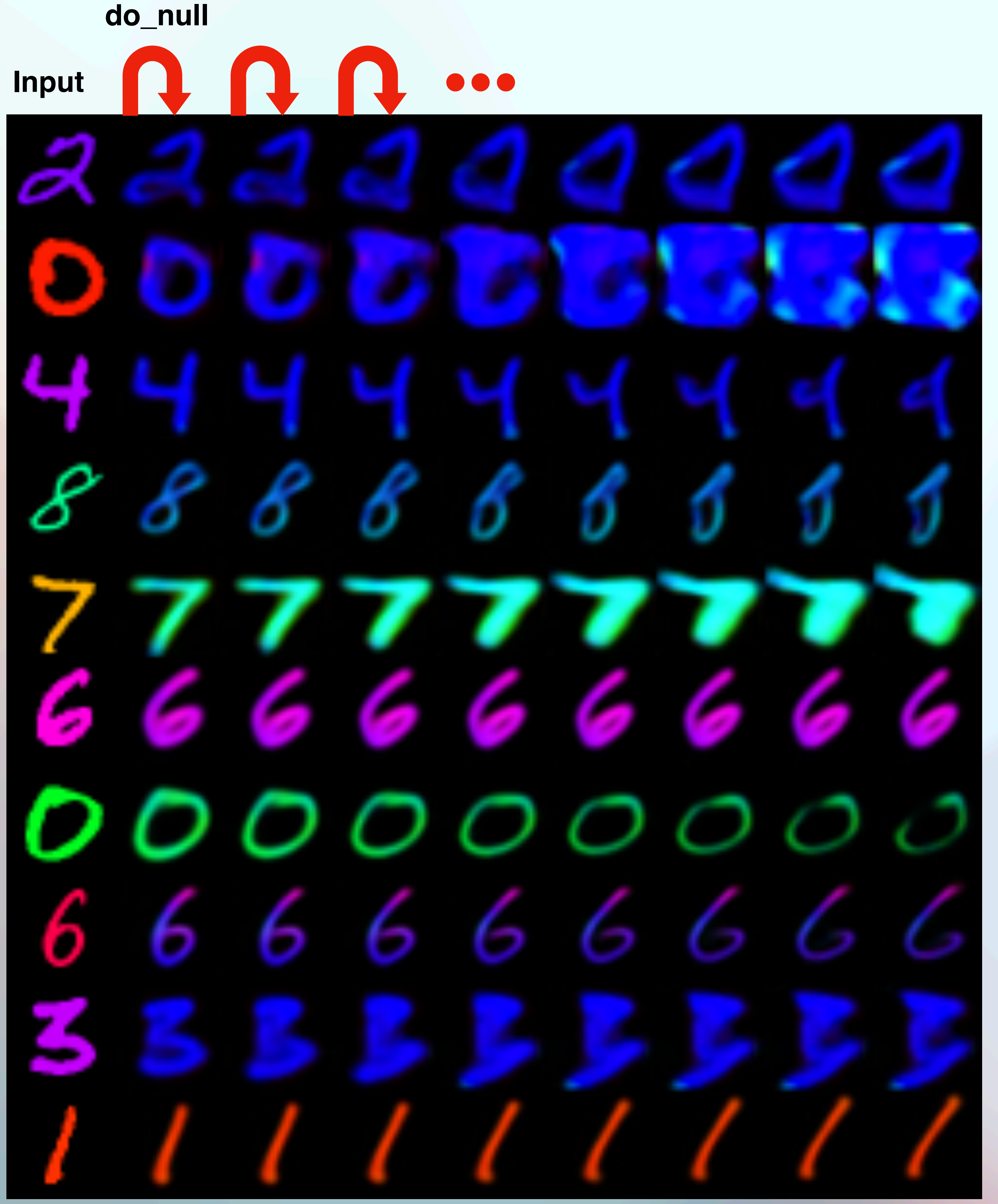
    1. A de-biased model: Normal VAE on the confounded scenario w/ full support and a simulated intervention;

    2. A biased model: Normal VAE on the confounded scenario w/o full support and no simulated intervention.

Colour MNIST composition



De-biased model



Biased model

Colour MNIST digit effectiveness

De-biased model

Biased model

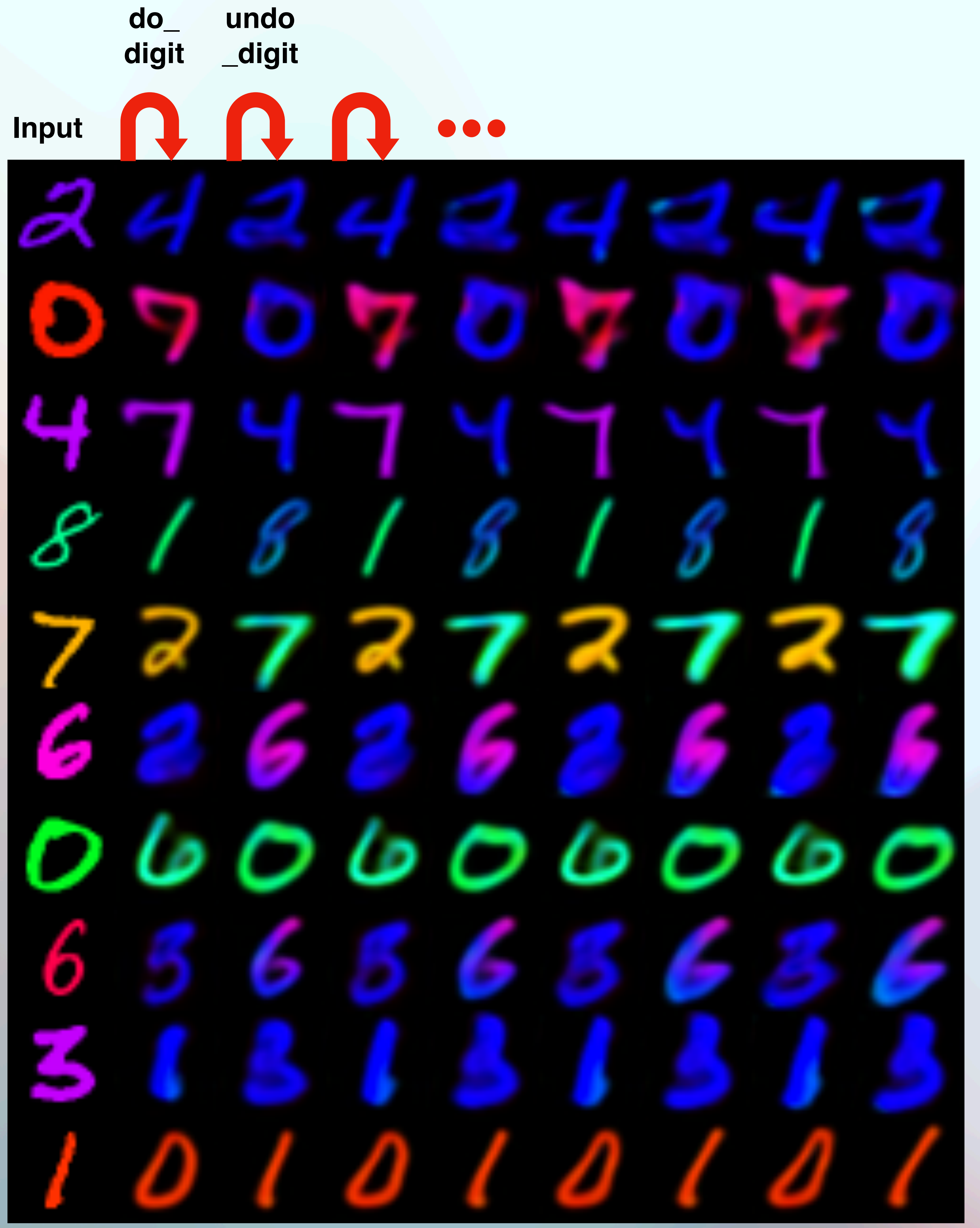
Colour MNIST hue effectiveness

De-biased model

Biased model

Colour MNIST digit reversibility

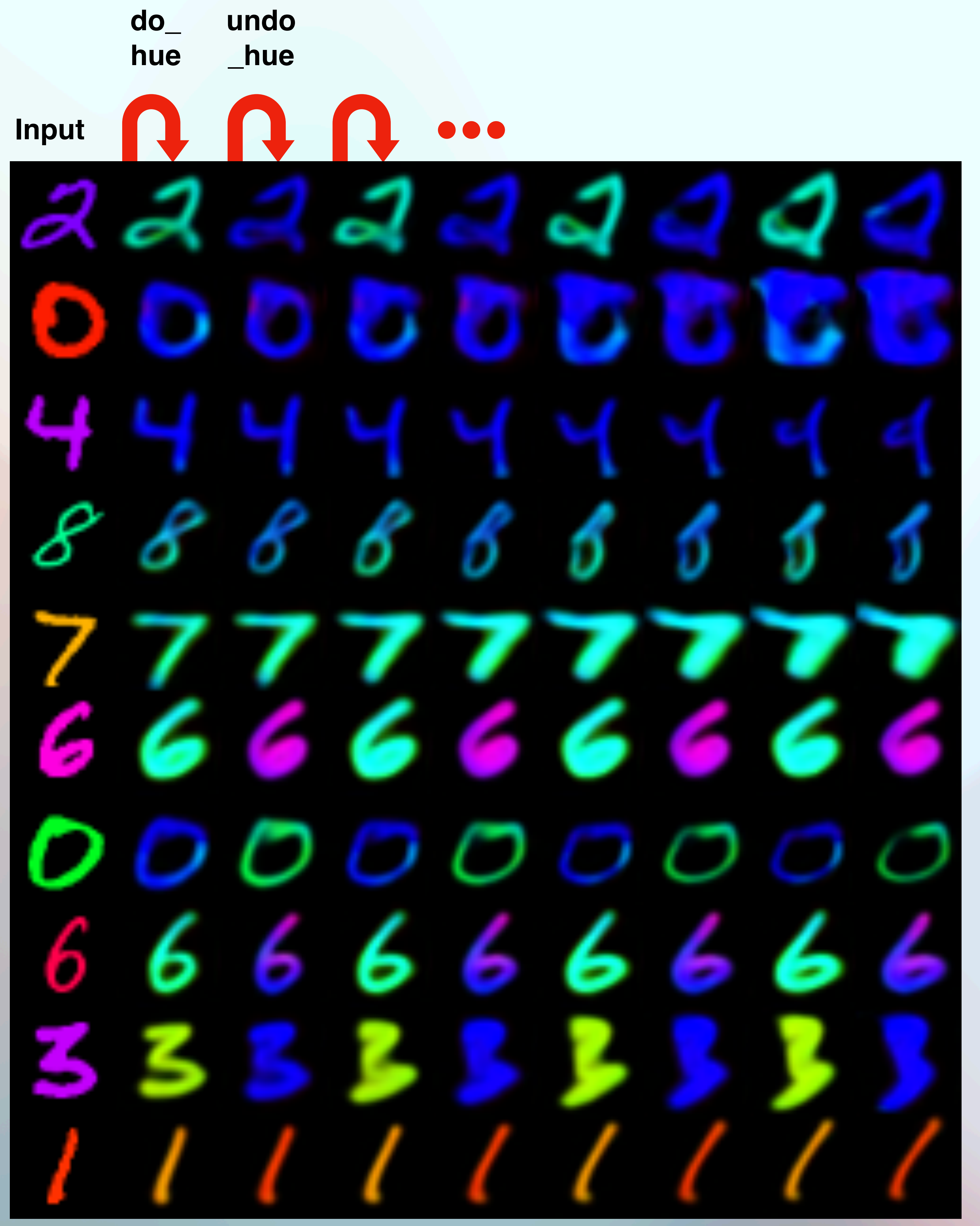
De-biased model



Biased model

Colour MNIST hue reversibility

De-biased model

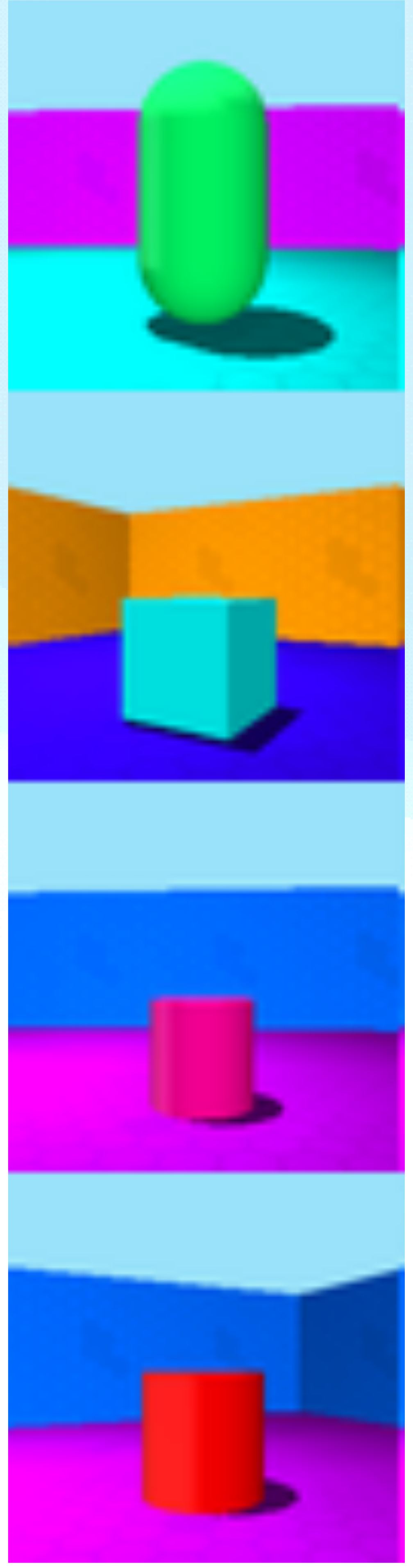


Biased model

Colour MNIST full results

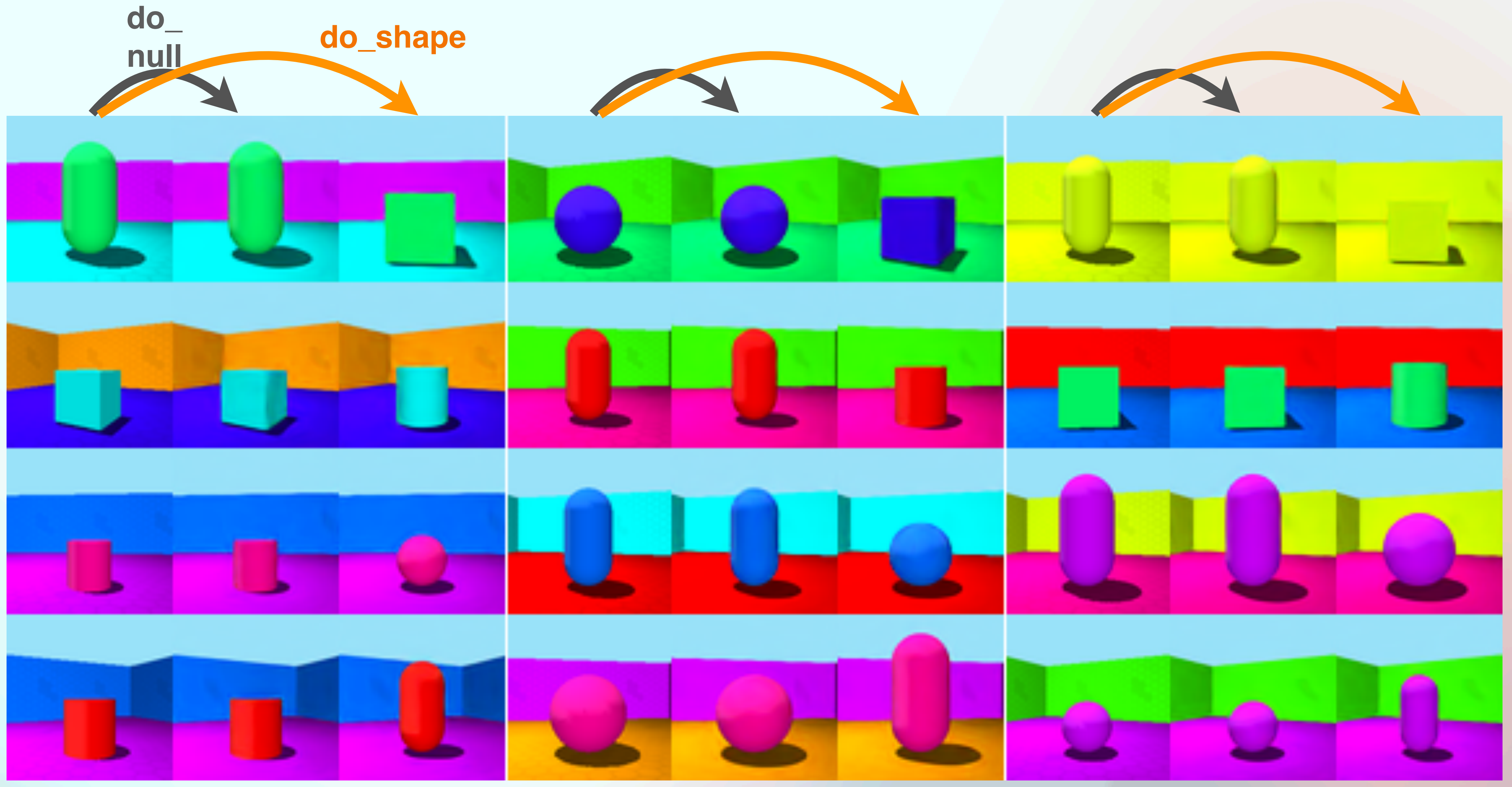
dataset	inter-ven-tion	model	null-intervention composition $l_1^{(1)} \downarrow$	digit intervention effectiveness $\text{acc}_{\text{digit}}(\%) \uparrow$	digit intervention effectiveness $\text{ae}_{\text{hue}}(\%) \downarrow$	digit intervention reversibility $l_1^{(1)} \downarrow$	hue intervention effectiveness $\text{acc}_{\text{digit}}(\%) \uparrow$	hue intervention effectiveness $\text{ae}_{\text{hue}}(\%) \downarrow$	hue intervention reversibility $l_1^{(1)} \downarrow$
un-con-found-ed	-	Identity	0.00	10.50	1.38	0.00	99.18	32.98	0.00
	-	VAE w/o encoder	19.04 (0.09)	97.08 (0.25)	1.32 (0.05)	19.04 (0.09)	97.24 (0.26)	1.32 (0.06)	19.04 (0.09)
	-	Bernoulli VAE β=1	5.98 (0.06)	98.68 (0.13)	1.29 (0.04)	7.67 (0.06)	99.45 (0.09)	1.26 (0.05)	7.24 (0.05)
	-	Bernoulli VAE β=2	6.86 (0.07)	99.52 (0.07)	1.33 (0.15)	9.10 (0.12)	99.60 (0.04)	1.32 (0.15)	8.62 (0.11)
	-	Normal VAE β=5	6.26 (0.29)	97.24 (0.26)	1.52 (0.28)	8.07 (0.26)	99.38 (0.06)	1.47 (0.27)	7.51 (0.32)
	-	GAN	4.92 (0.05)	94.28 (1.01)	1.60 (0.22)	9.22 (0.27)	98.98 (0.05)	1.55 (0.23)	5.60 (0.03)
con-found-ed w/o full support	no	Bernoulli VAE β=1	9.20 (1.31)	97.12 (1.05)	10.74 (4.77)	11.42 (1.49)	98.89 (0.16)	11.60 (6.14)	11.11 (1.61)
		Bernoulli VAE β=2	10.84 (0.45)	98.94 (0.17)	10.36 (1.39)	12.82 (0.45)	99.17 (0.05)	10.07 (1.39)	12.52 (0.41)
		Normal VAE β=5	11.21 (0.63)	94.74 (0.51)	14.17 (2.63)	13.32 (0.62)	98.81 (0.22)	14.27 (3.03)	12.69 (0.59)
	yes	Bernoulli VAE β=1	8.63 (0.50)	96.94 (0.26)	6.38 (1.58)	11.10 (0.75)	98.88 (0.25)	7.02 (1.96)	10.79 (0.75)
		Bernoulli VAE β=2	9.85 (0.33)	95.76 (1.63)	6.44 (1.24)	12.10 (0.39)	95.77 (1.56)	6.44 (1.37)	11.86 (0.29)
		Normal VAE β=5	9.32 (1.41)	95.35 (0.71)	7.54 (1.99)	11.29 (1.39)	98.79 (0.28)	7.30 (2.03)	10.85 (1.36)
con-found-ed w/ full support	no	Bernoulli VAE β=1	6.68 (0.27)	96.62 (2.09)	8.52 (6.93)	8.89 (0.70)	99.20 (0.10)	12.15 (11.69)	8.45 (0.69)
		Bernoulli VAE β=2	7.56 (0.10)	99.36 (0.16)	2.70 (0.12)	9.67 (0.06)	99.47 (0.06)	2.54 (0.12)	9.32 (0.09)
		Normal VAE β=5	6.72 (0.30)	95.53 (0.28)	3.88 (1.12)	9.06 (0.68)	99.07 (0.04)	3.59 (1.20)	8.45 (0.67)
	yes*	GAN	6.05 (0.06)	95.17 (0.55)	1.95 (0.07)	11.18 (0.10)	99.18 (0.10)	1.73 (0.11)	7.79 (0.10)
	yes	Bernoulli VAE β=1	6.67 (0.10)	99.07 (0.15)	2.31 (0.24)	8.42 (0.16)	99.37 (0.13)	3.08 (1.08)	8.40 (0.48)
		Bernoulli VAE β=2	7.84 (0.09)	99.63 (0.03)	2.16 (0.06)	9.63 (0.08)	99.61 (0.06)	2.01 (0.10)	9.34 (0.09)
		Normal VAE β=5	6.51 (0.29)	97.75 (0.18)	3.05 (0.44)	8.35 (0.29)	99.31 (0.07)	2.73 (0.47)	7.83 (0.31)
		GAN	5.25 (0.06)	96.27 (0.26)	1.84 (0.11)	10.75 (0.34)	99.01 (0.06)	1.77 (0.14)	6.20 (0.04)

3D Shapes
Experiments and results

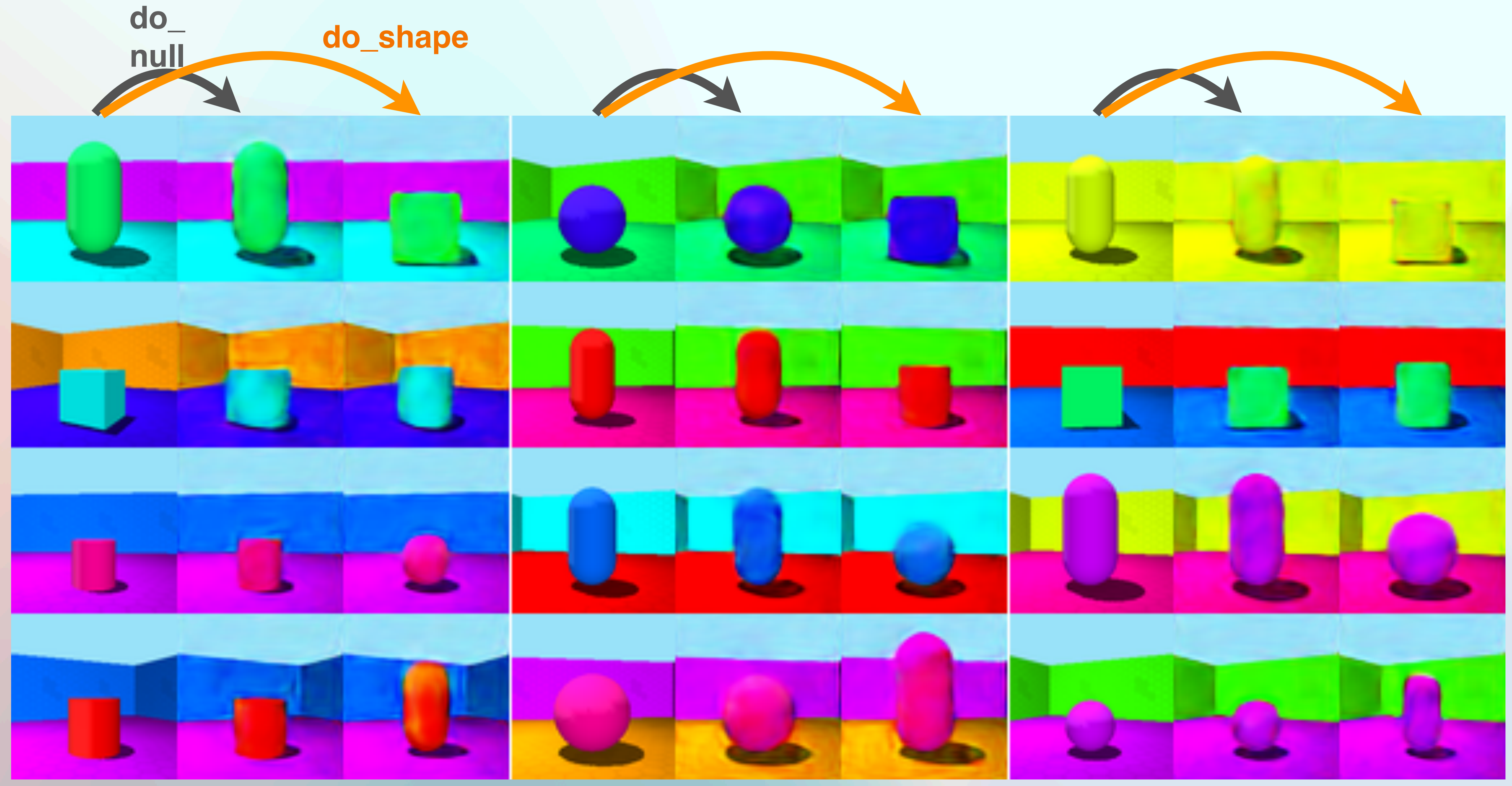


- Procedurally generated images from 6 parents:
  1. Object hue;
  2. Object shape;
  3. Object size;
  4. Object rotation angle;
  5. Wall hue;
  6. Floor hue.
- In theory, there is no exogenous noise, image is fully determined by parents.

3D Shapes (object shape)

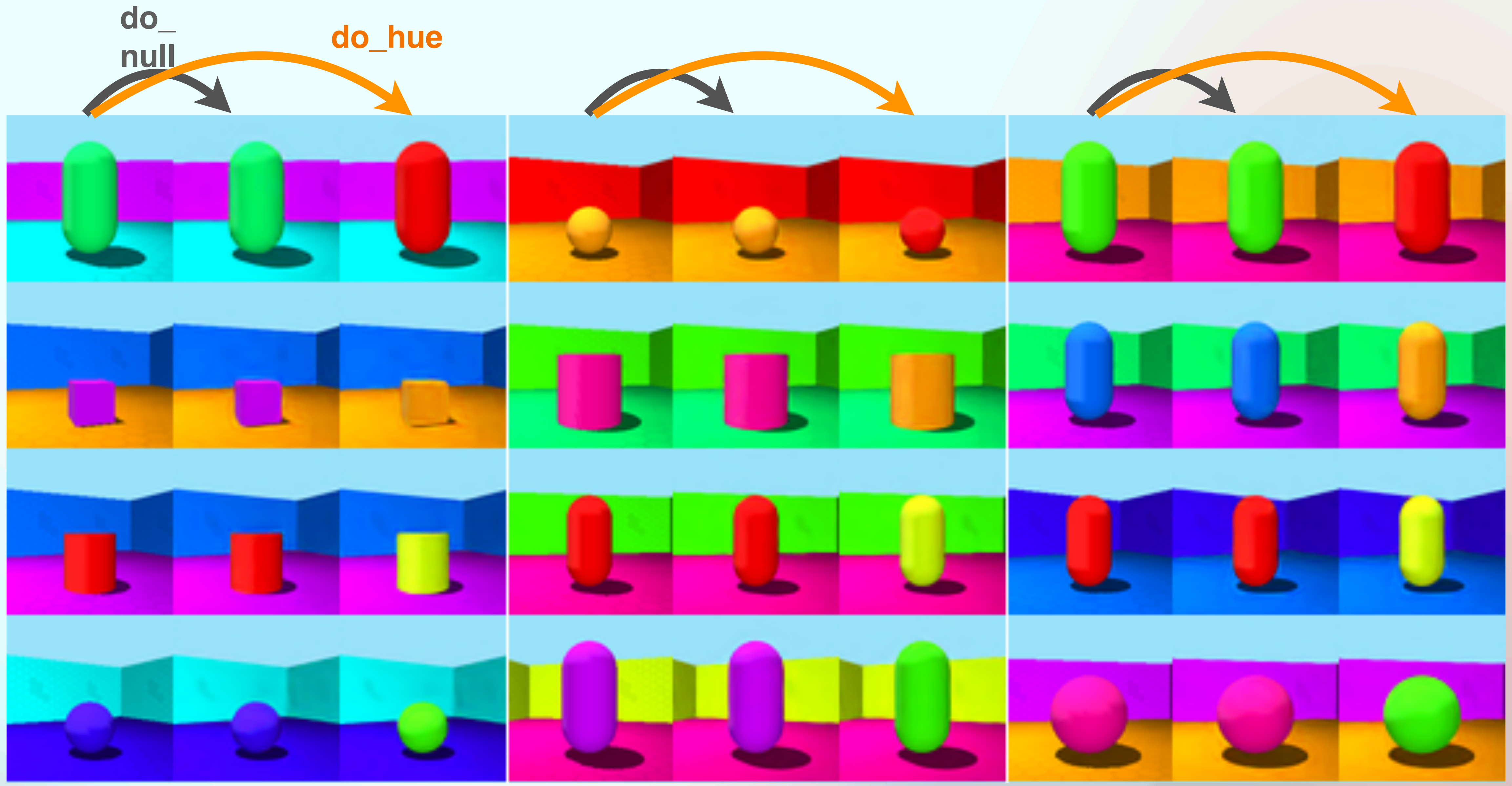


VAE

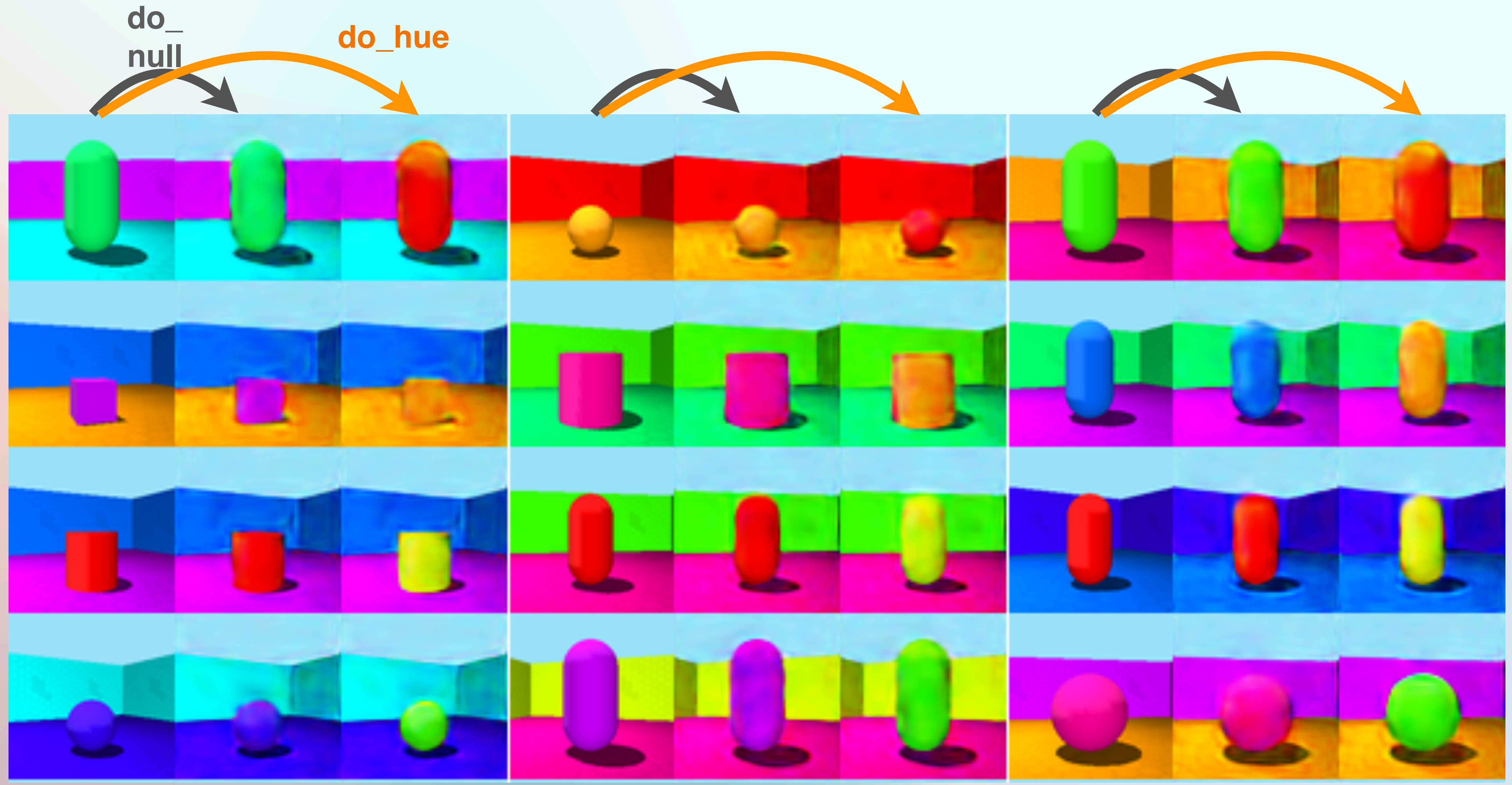


GAN

3D Shapes (object hue)



VAE



GAN

CELEB-A HQ

Experiments and results

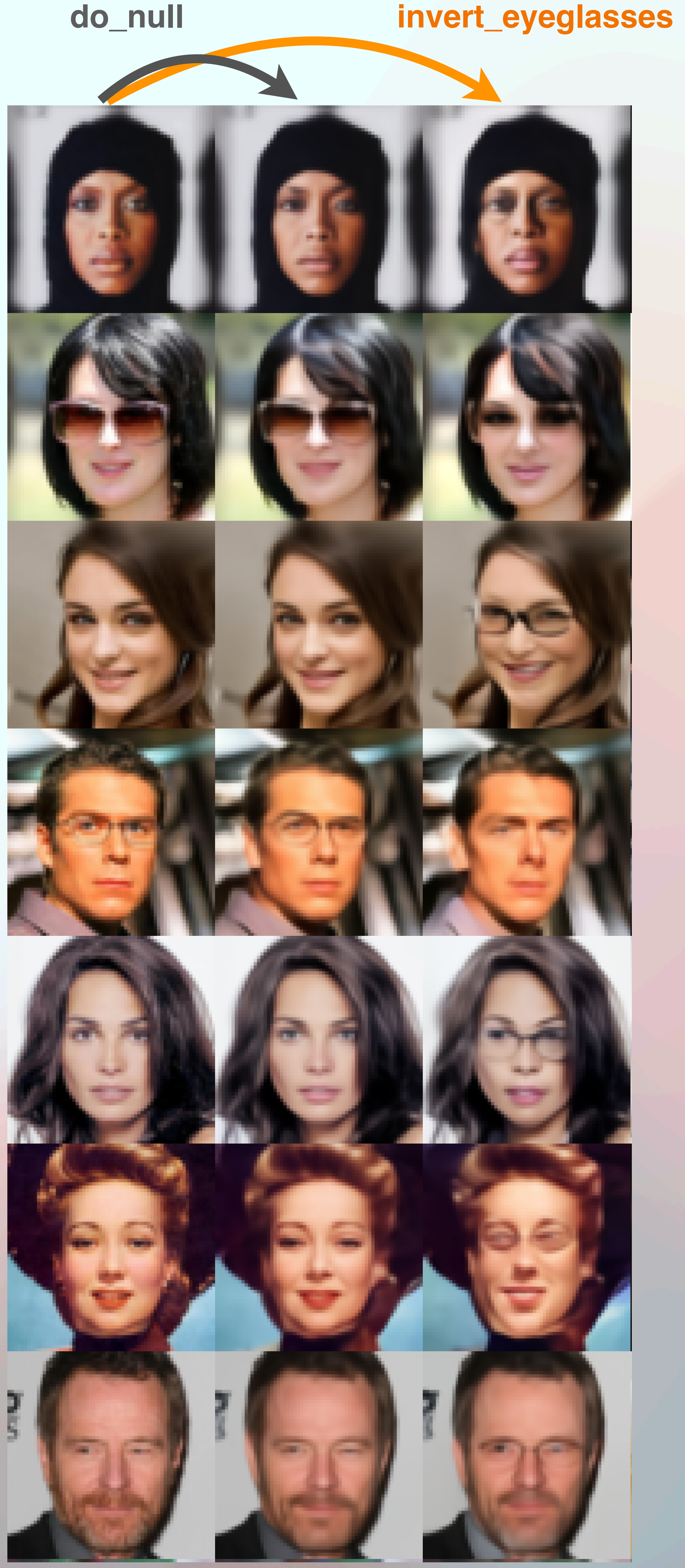
- Deep hierarchical VAE with 42 latent variables;
- Counterfactuals can be produced by abducting all latent variables or only a subset;
- Instead of abducting variables we can sample from the exogenous noise distribution (technically not a “real” counterfactual);
- We see a trade-off between obeying the counterfactual conditioning and preserving subject identity.

CELEB-A HQ smiling counterfactuals

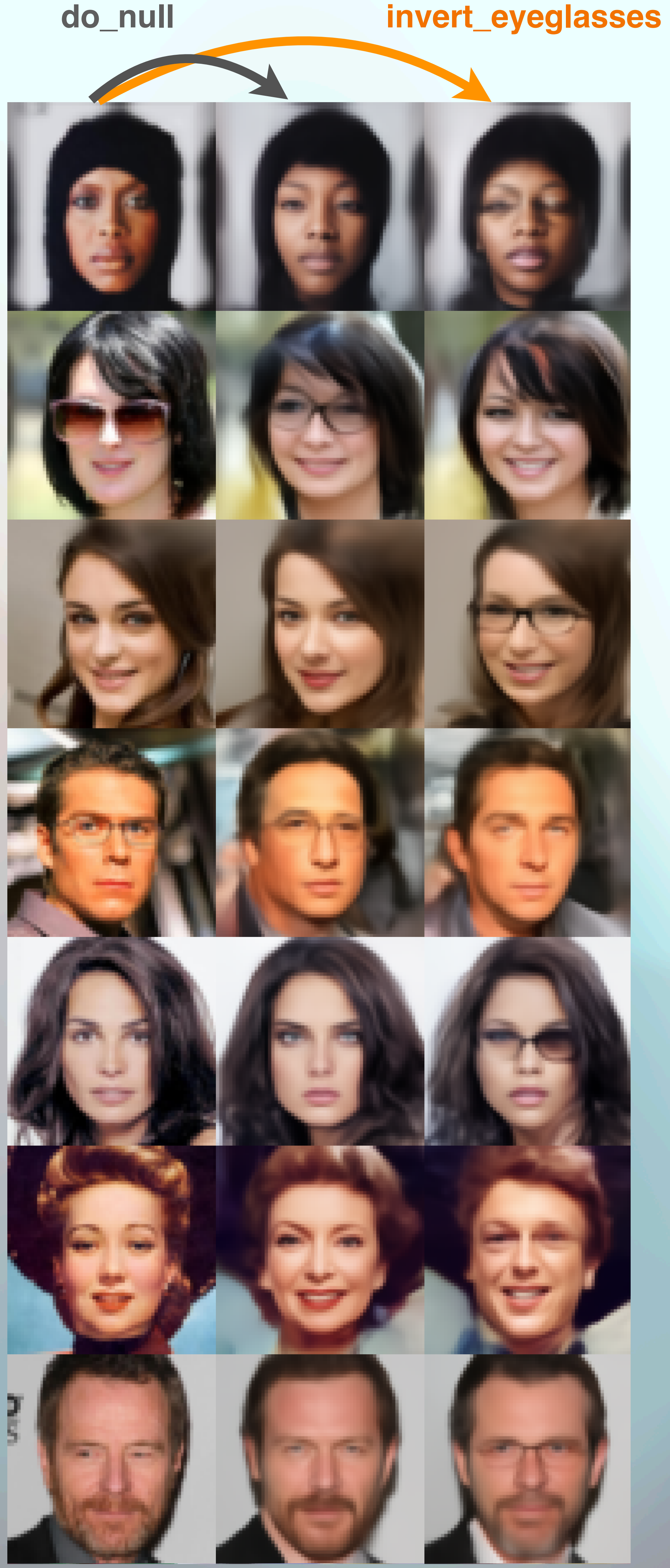
42 latents

8 latents

CELEB-A HQ eye-glasses counterfactuals



42 latents



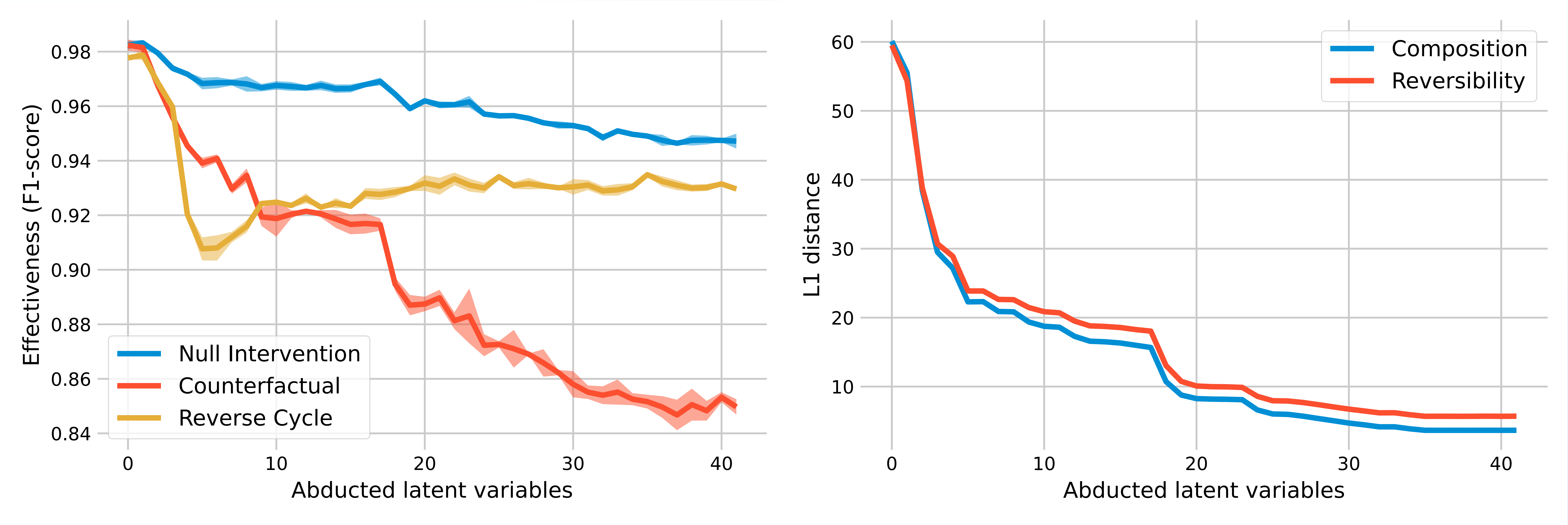
8 latents

CELEB-A HQ trade-off (1)

Experiments and results

- We see a trade-off between obeying the counterfactual conditioning and preserving subject identity;
- In other words, there is a trade-off between composition and effectiveness for these two models.

CELEB-A HQ trade-off (2)



Conclusion

- The axioms of composition, effectiveness and reversibility provide a theoretical grounded manner of evaluating and comparing counterfactual image models;
- The axioms lead us to a set of soundness metrics which allow to compare approximate causal models with each other and against an unavailable ideal model;

Thank you

Questions?