Evaluating the Disentanglement of Deep Generative Models with Manifold Topology

Sharon Zhou, Eric Zelikman, Fred Lu,
Andrew Ng, Gunnar Carlsson, Stefano Ermon

Computer Science & Math Departments, Stanford University

ICLR 2021

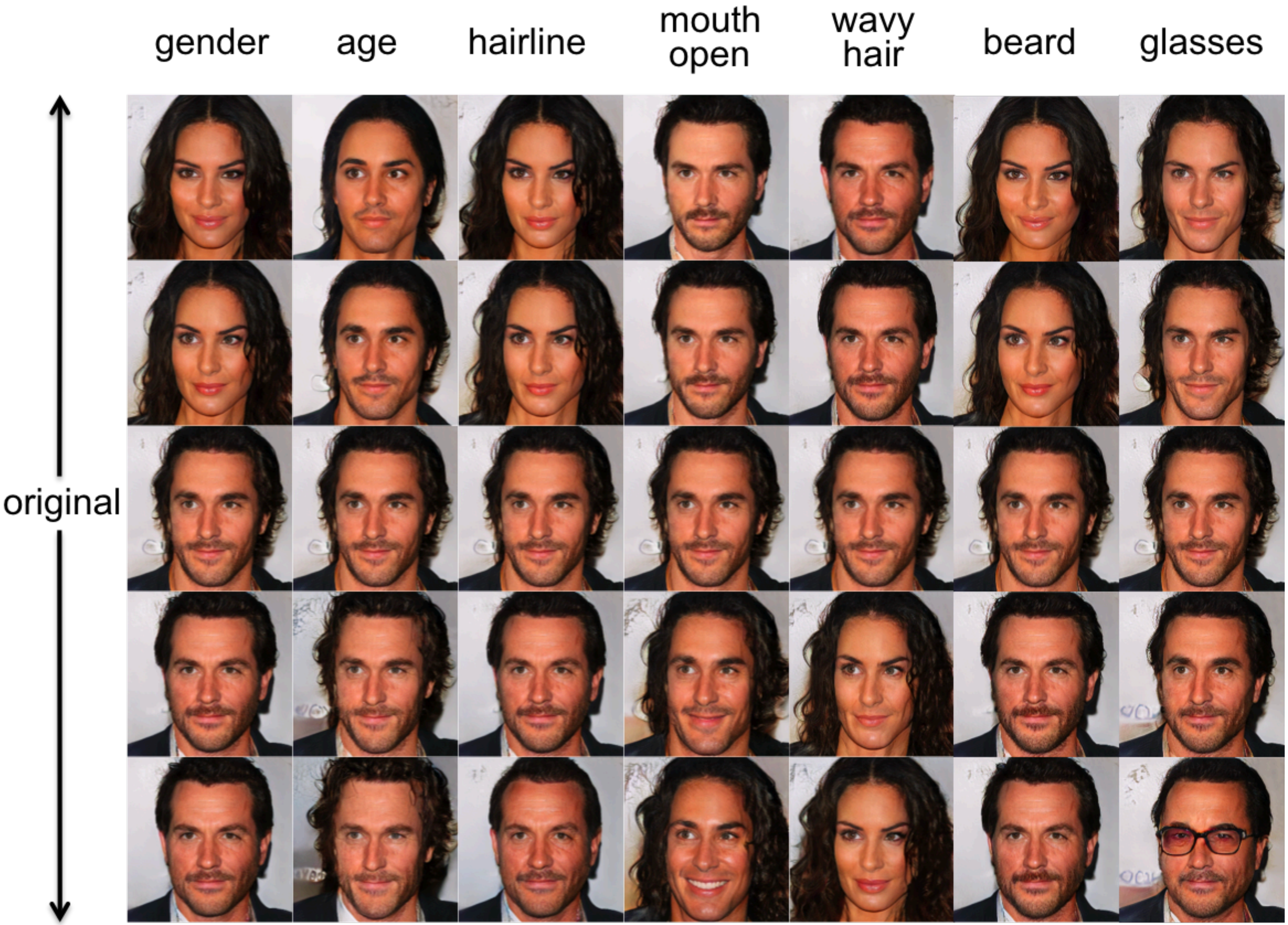


Photo credit: https://blog.insightdatascience.com/generating-custom-photo-realistic-faces-using-ai-d170b1b59255

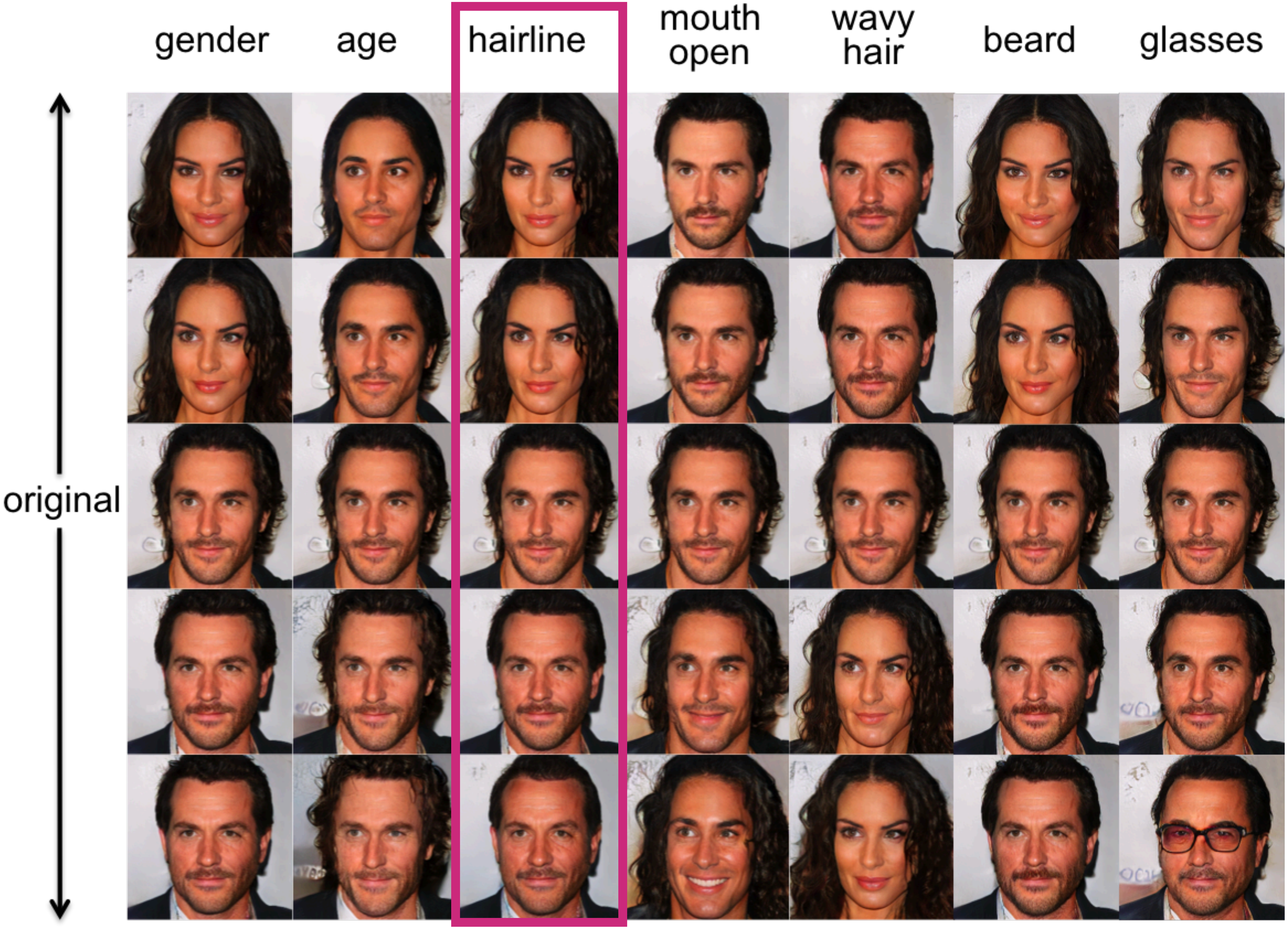


Photo credit: https://blog.insightdatascience.com/generating-custom-photo-realistic-faces-using-ai-d170b1b59255

Good Representations Disentangle the Explanatory Factors of Variation*

Glasses

Beard

Hair

Eyes

Age

* Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

How would you describe all the major features of *dogs*?

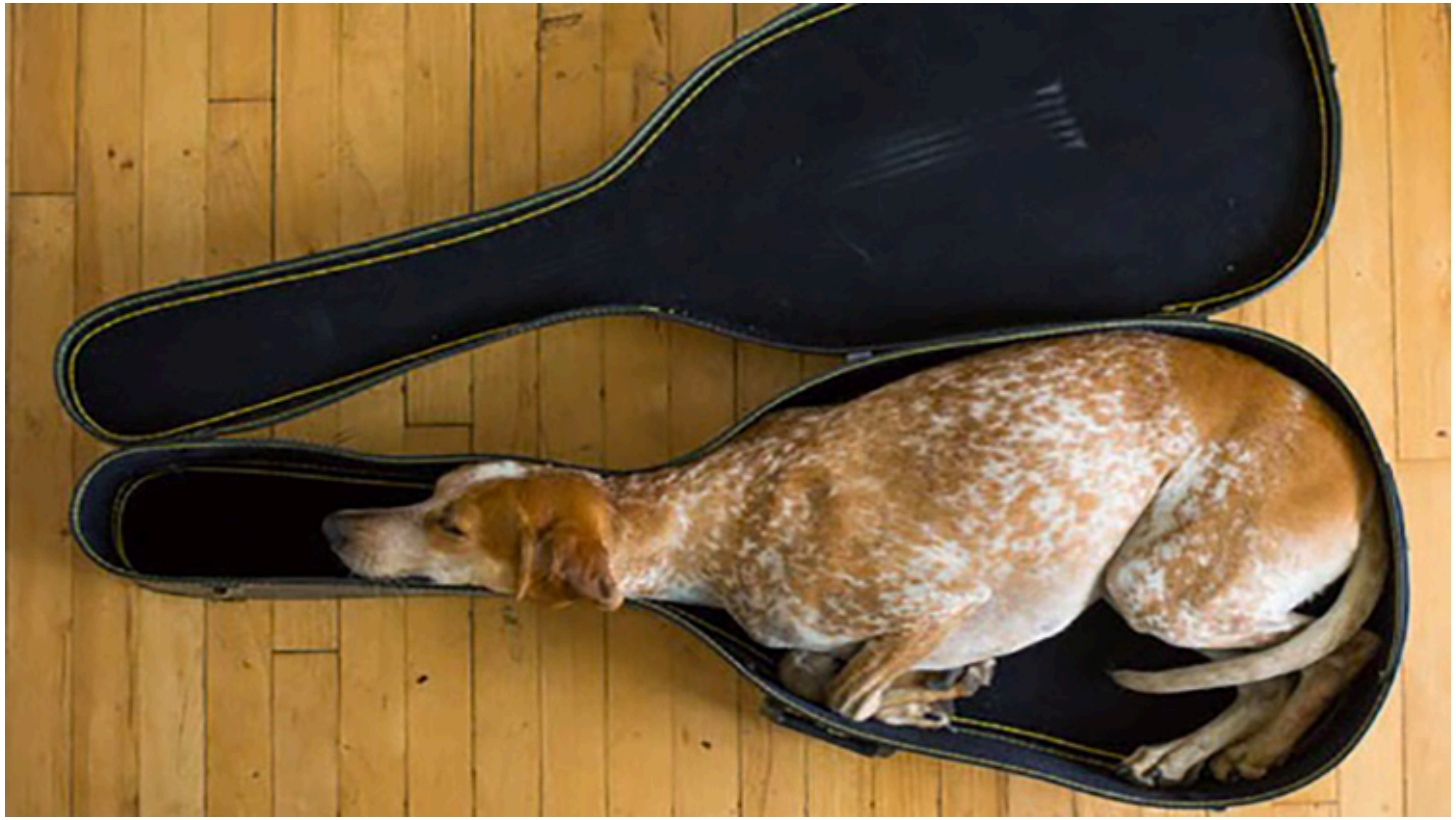
Size

Fur type

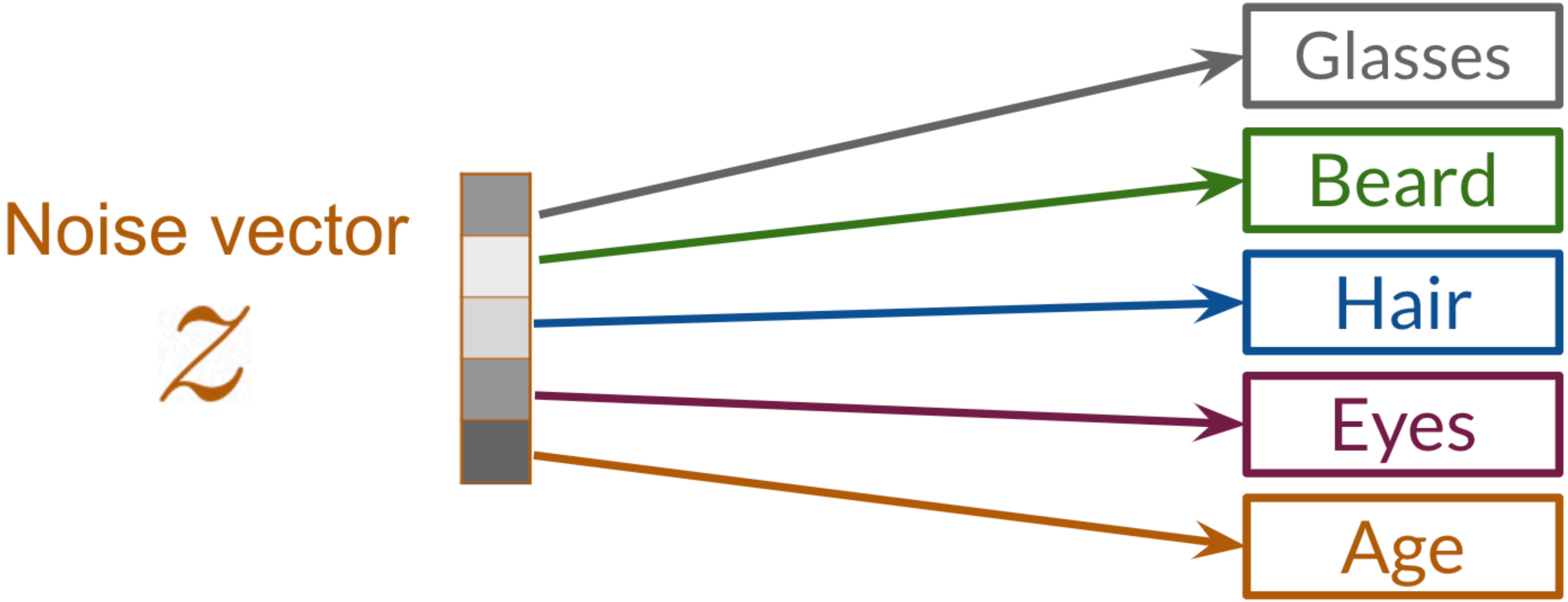
Nose color

Ear floppiness

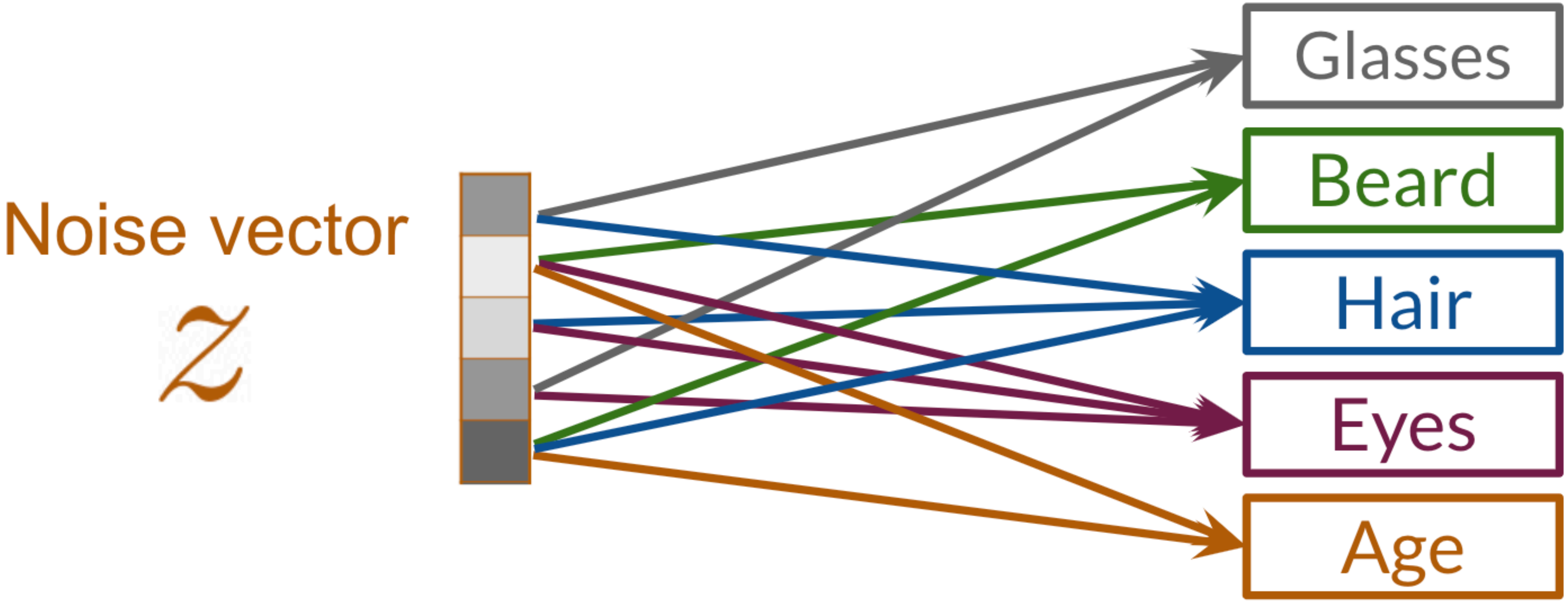
Tongue droopiness



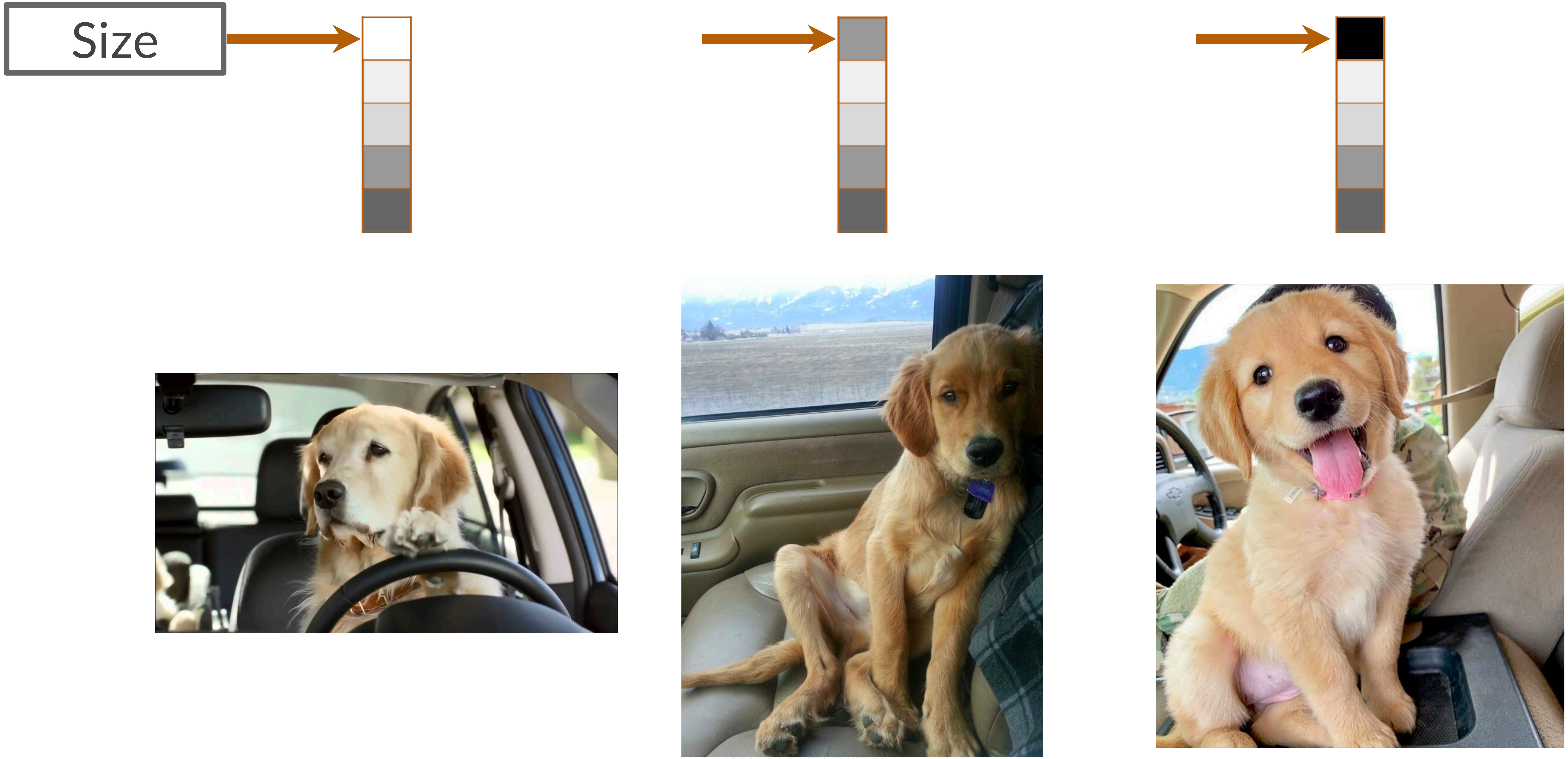
Z-Space Disentanglement



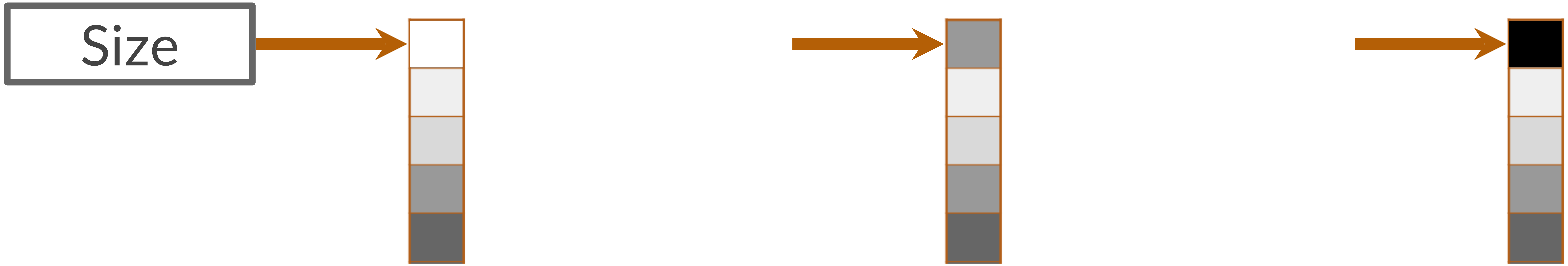
Z-Space Entanglement

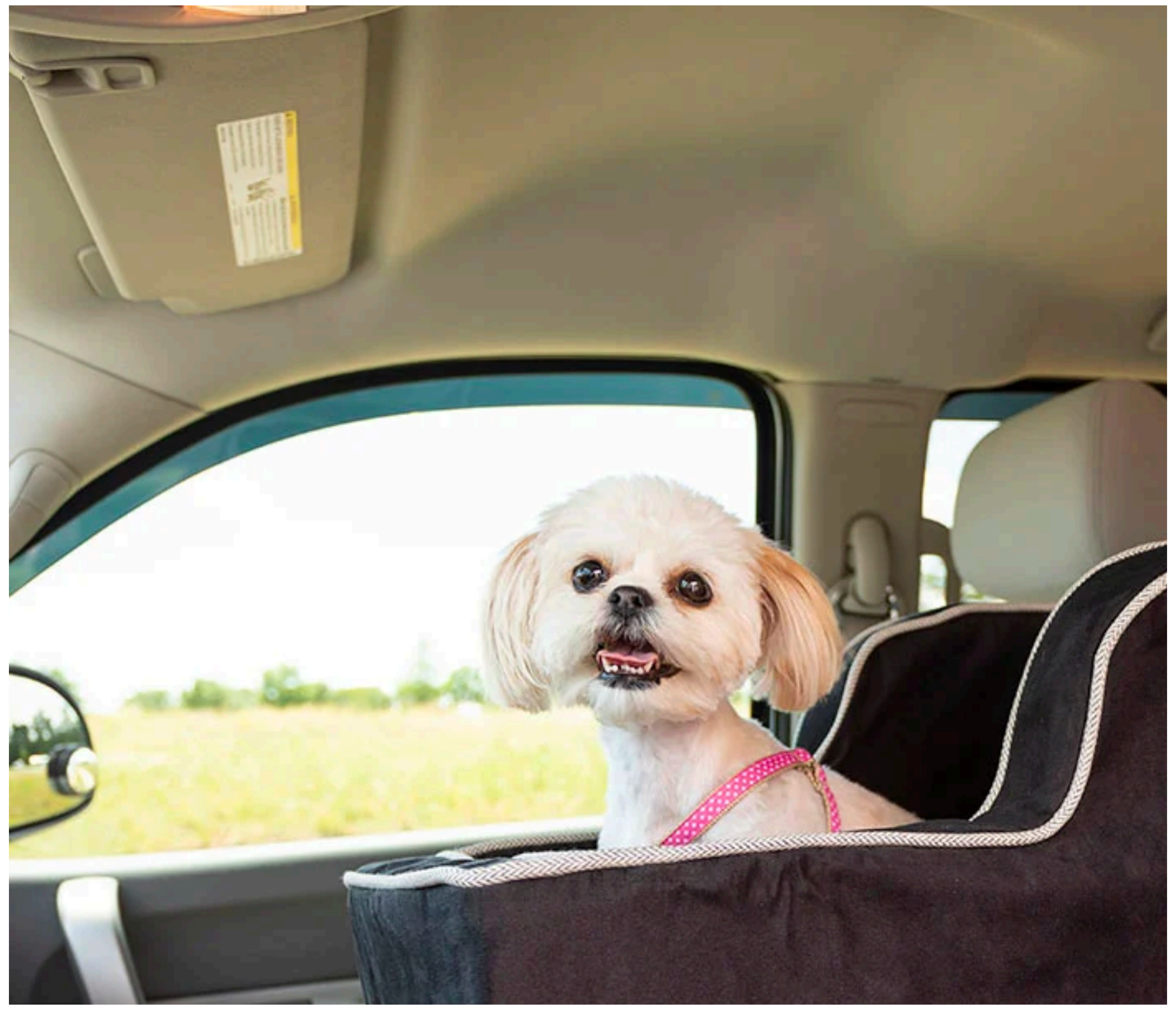


Disentangled latent dimension

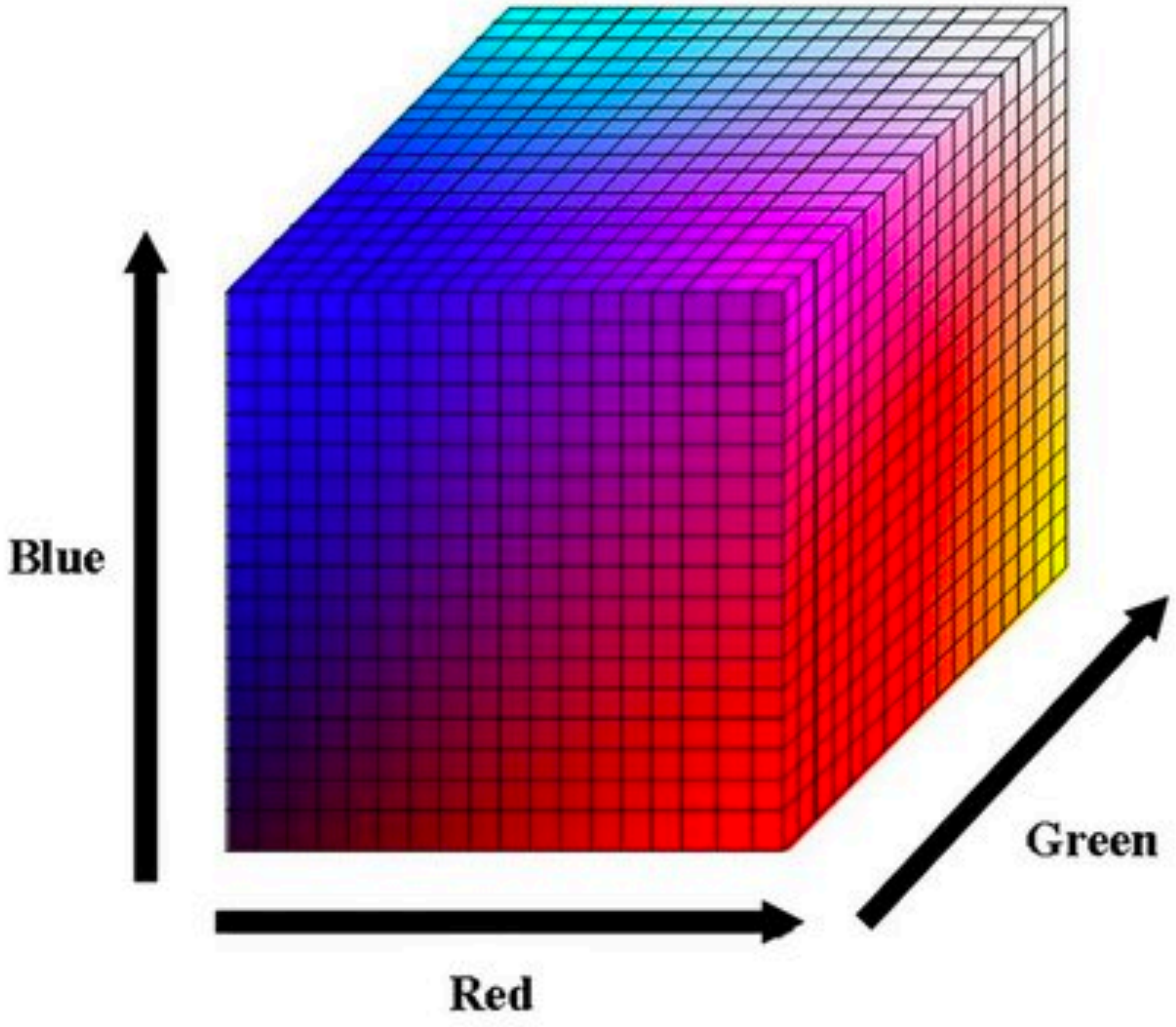


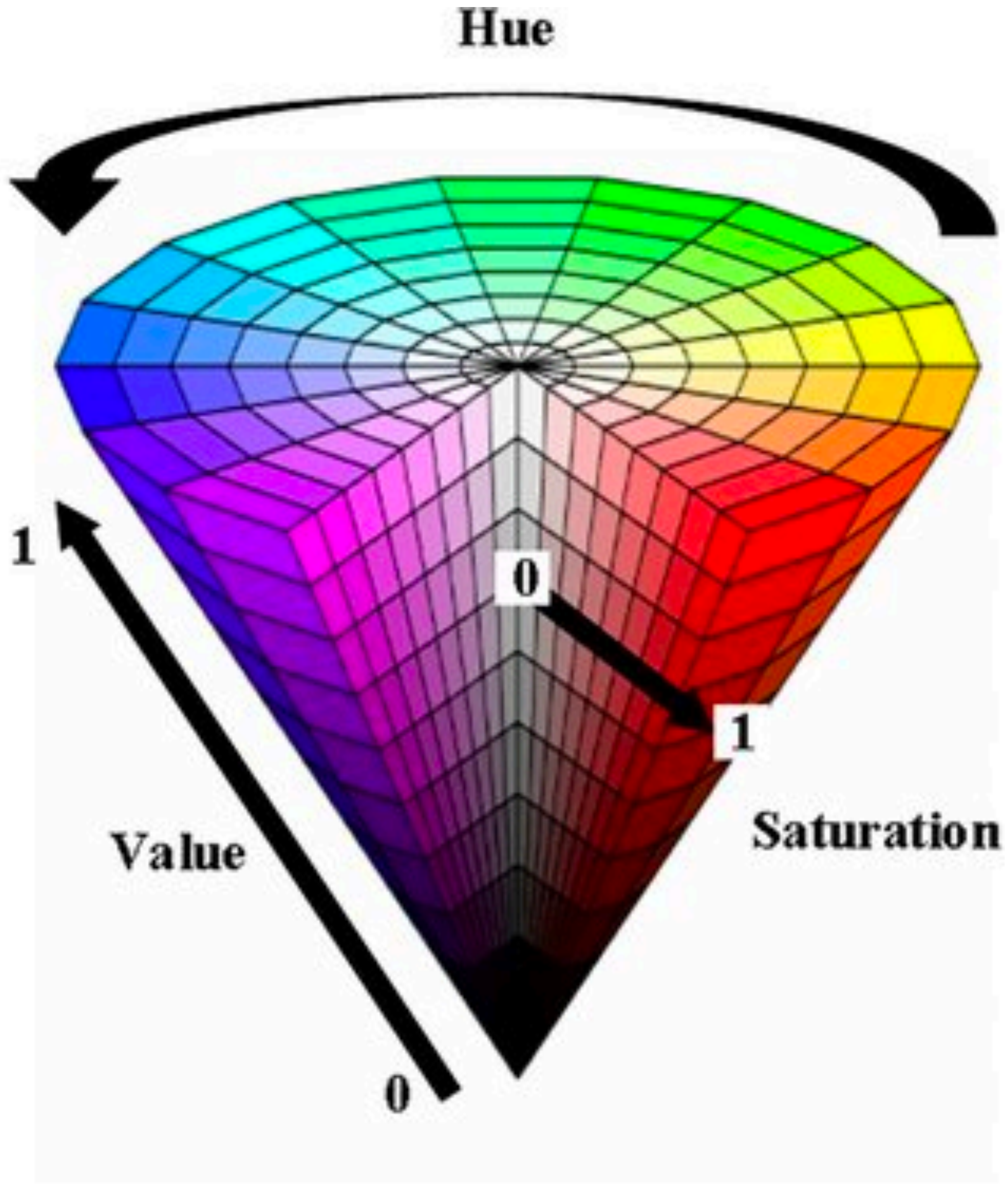
Entangled latent dimension





Multiple Factorizations Are Possible





Chen, Rui, Meiling Wang, and Yi Lai. "Analysis of the role and robustness of artificial intelligence in commodity image recognition under deep learning neural network." *Plos one* 15.7 (2020): e0235783.

dSprites Disentanglement Dataset

1. Shape (square, ellipse, heart)
2. Scale (size)
3. Orientation (rotation)
4. X Position
5. Y Position

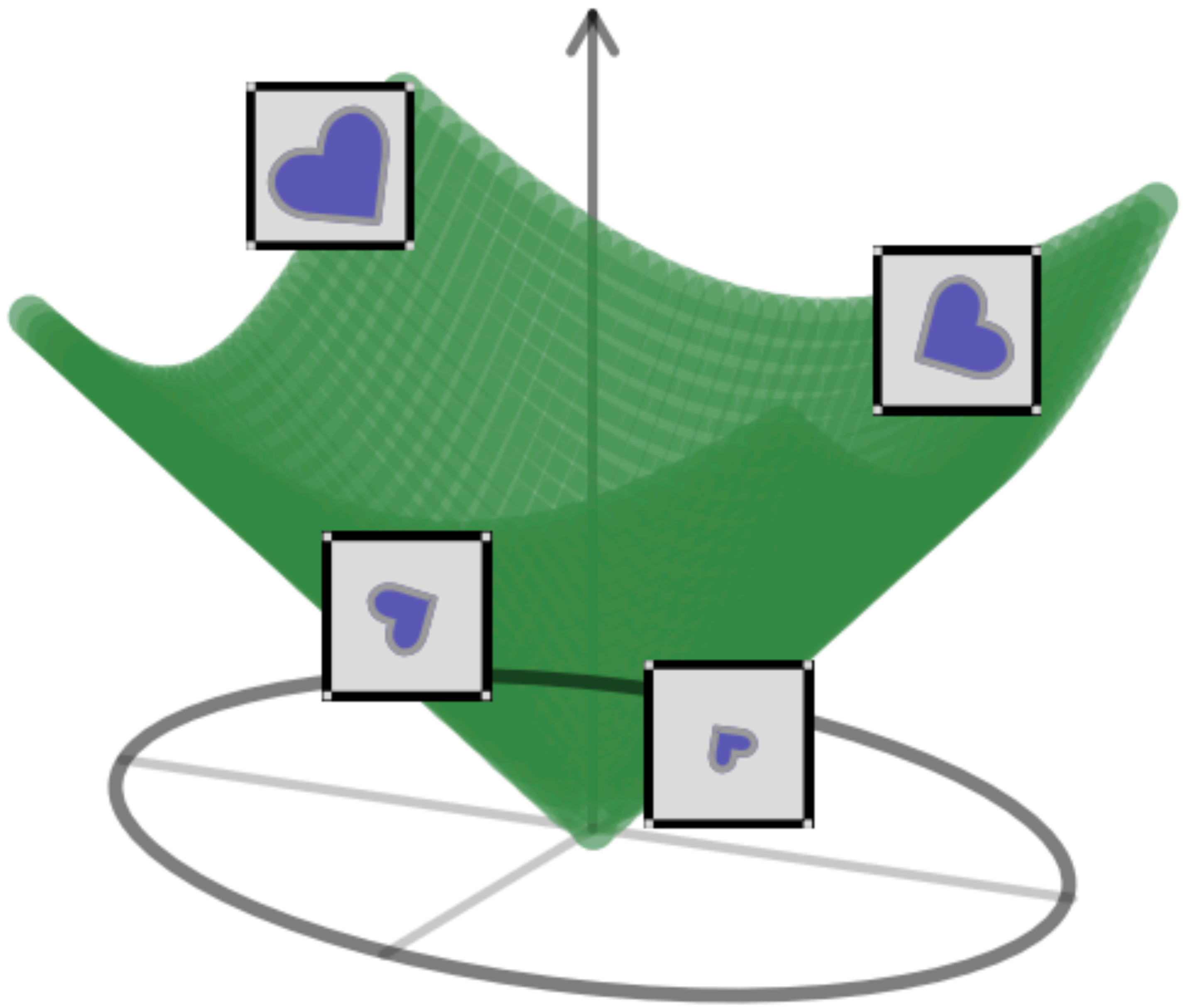
L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. *dsprites*: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Prior methods

1. **Supervision** required for a specific factorization
2. Tuned to a **specific dataset**, e.g. custom preprocessing on face images
3. Depends on the **architecture** specifically with an **external model**, e.g. encoder and/or classifier

Ours

1. **Unsupervised** and supervised variants both available
2. Procedure can be **applied across datasets** — and architectures, as above
3. Uses an **intrinsic property** of a generative model, without reliance on external models or custom architectures



Scale
Rotation

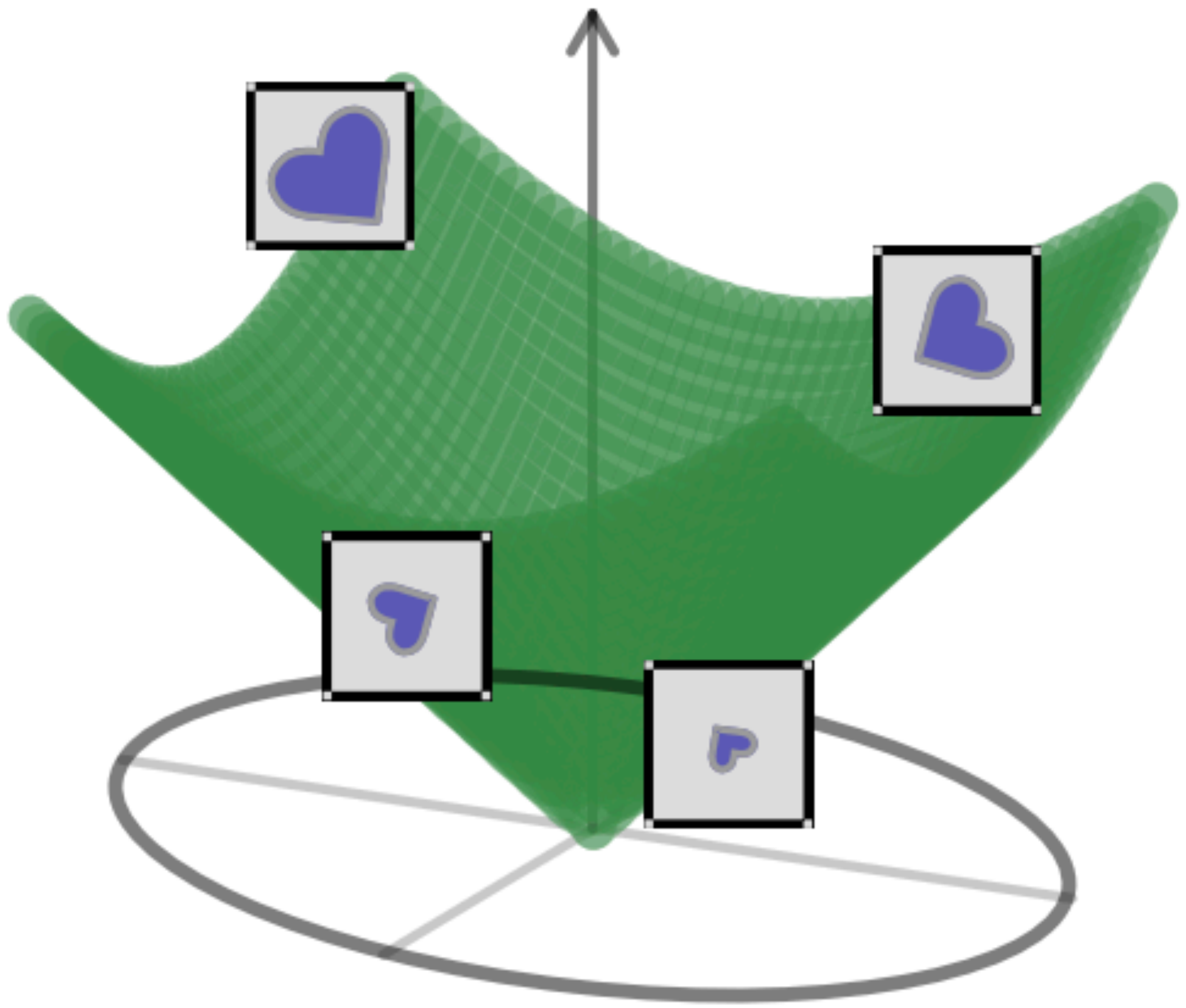
Scale
Rotation

Scale
Rotation

Scale
Rotation

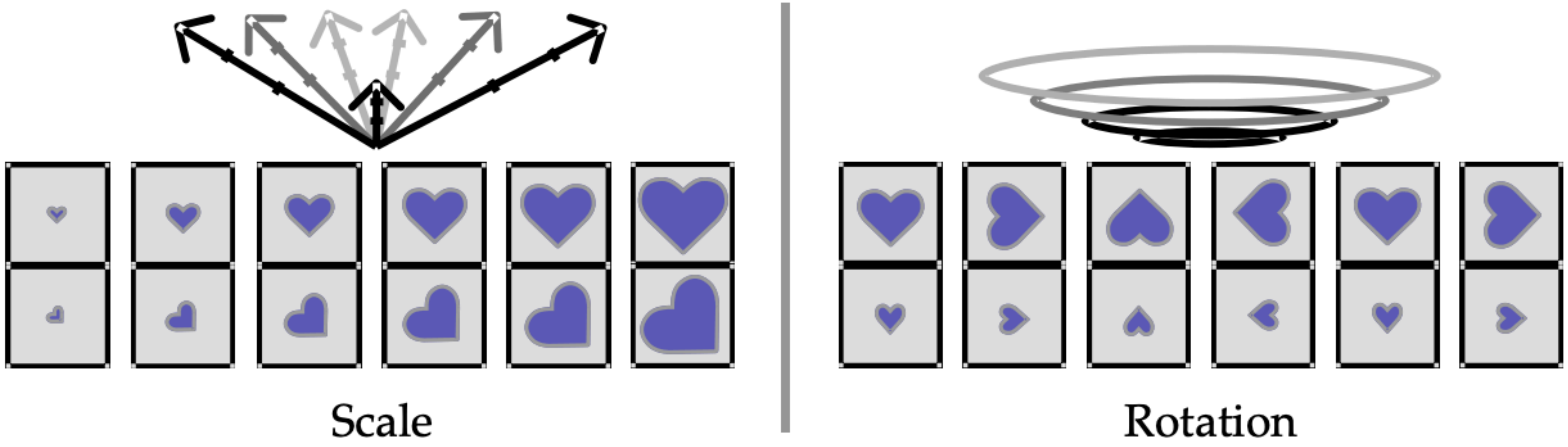
Scale
Rotation

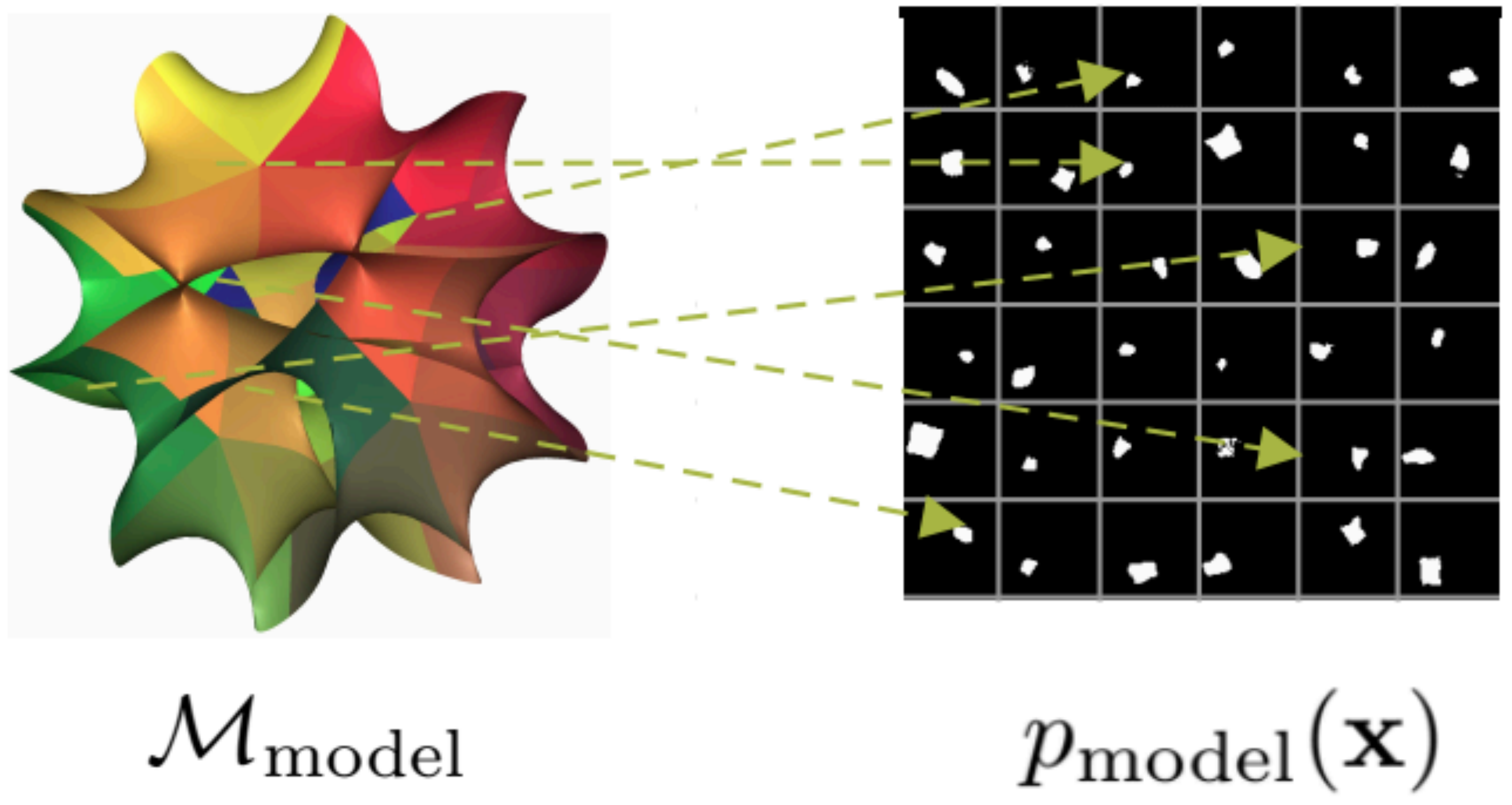
Scale
Rotation



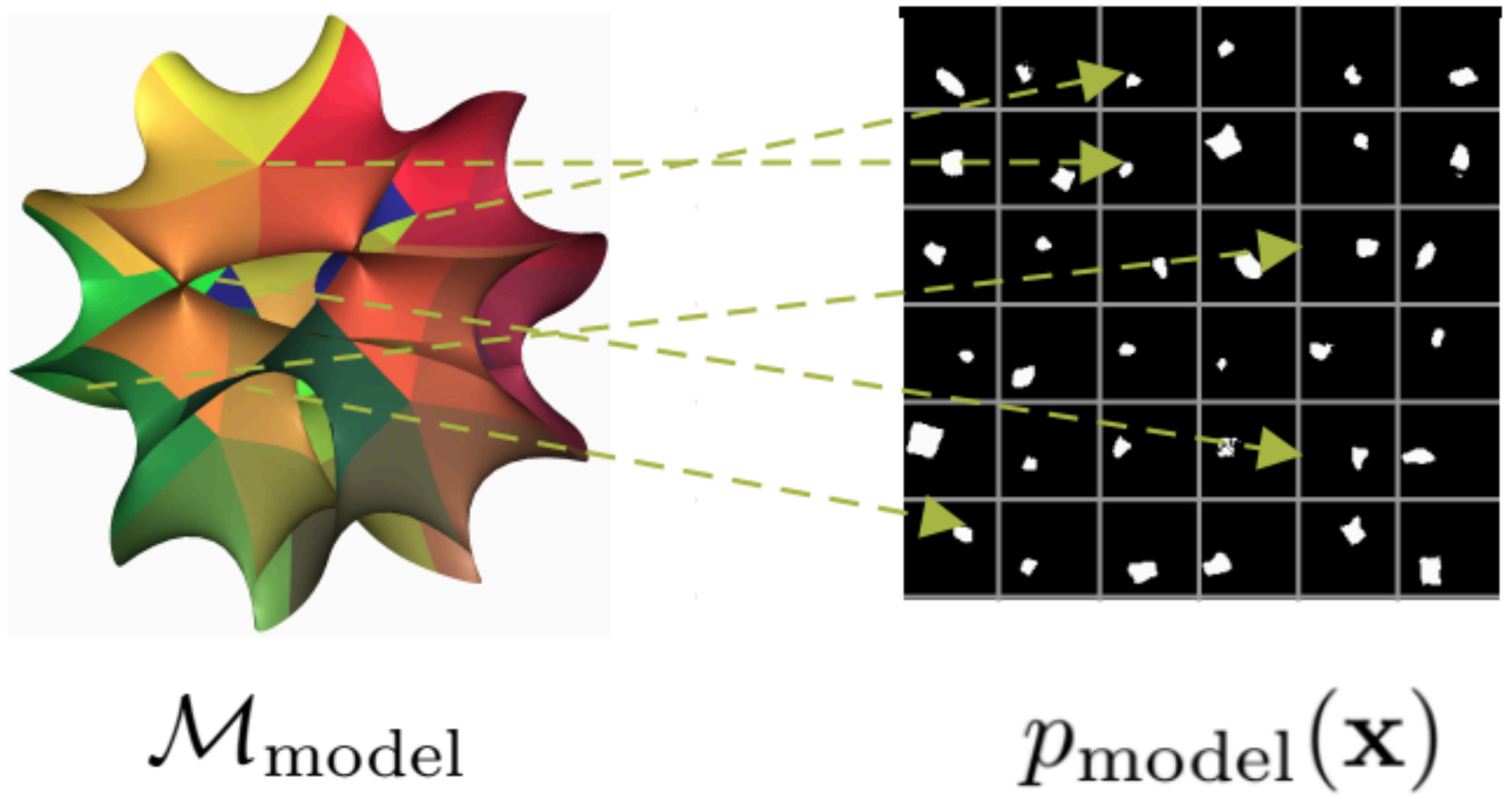
No holes

1D hole



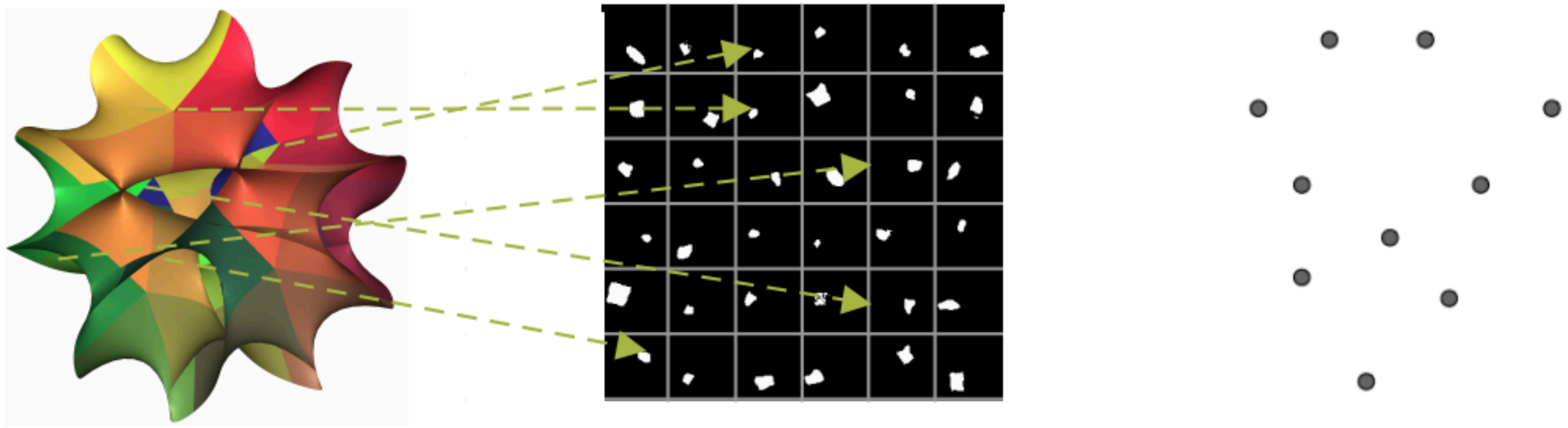


Manifold subfigure (a) is inspired by Hanson (1994)
Data points and corresponding simplices subfigures (b)-(c) by Khrulkov and Oseledets (2018)



Manifold subfigure (a) is inspired by Hanson (1994)

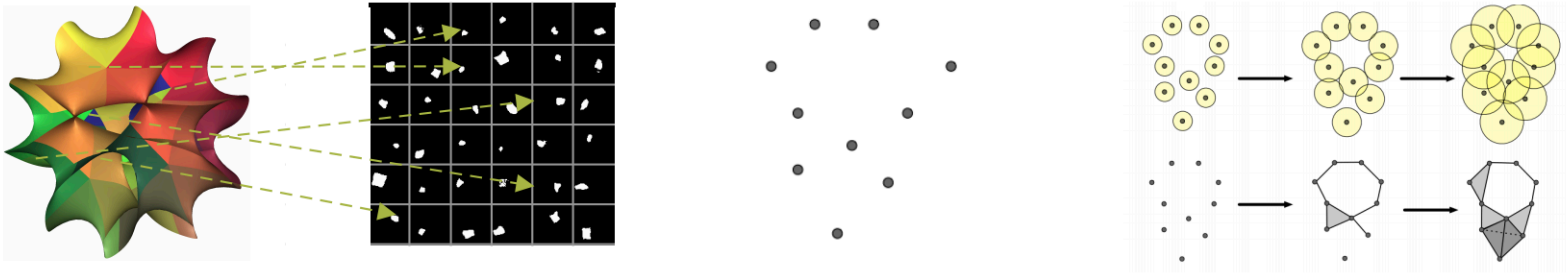
Data points and corresponding simplices subfigures (b)-(c) by Khrulkov and Oseledets (2018)



$\mathcal{M}_{\text{model}}$ $p_{\text{model}}(\mathbf{x})$ $X \sim p_{\text{model}}(\mathbf{x})$

Manifold subfigure (a) is inspired by Hanson (1994)

Data points and corresponding simplices subfigures (b)-(c) by Khrulkov and Oseledets (2018)



$\mathcal{M}_{\text{model}}$ $p_{\text{model}}(\mathbf{x})$ $X \sim p_{\text{model}}(\mathbf{x})$ Simplices, increasing proximity

Manifold subfigure (a) is inspired by Hanson (1994)
Data points and corresponding simplices subfigures (b)-(c) by Khrulkov and Oseledets (2018)

0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60

$\mathcal{M}_{\text{model}}$ $p_{\text{model}}(\mathbf{x})$ $X \sim p_{\text{model}}(\mathbf{x})$ Simplices, increasing proximity Wasserstein RLT

Manifold subfigure (a) is inspired by Hanson (1994)
Data points and corresponding simplices subfigures (b)-(c) by Khrulkov and Oseledets (2018)

0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60

$\mathcal{M}_{\text{model}}$

$p_{\text{model}}(\mathbf{x})$

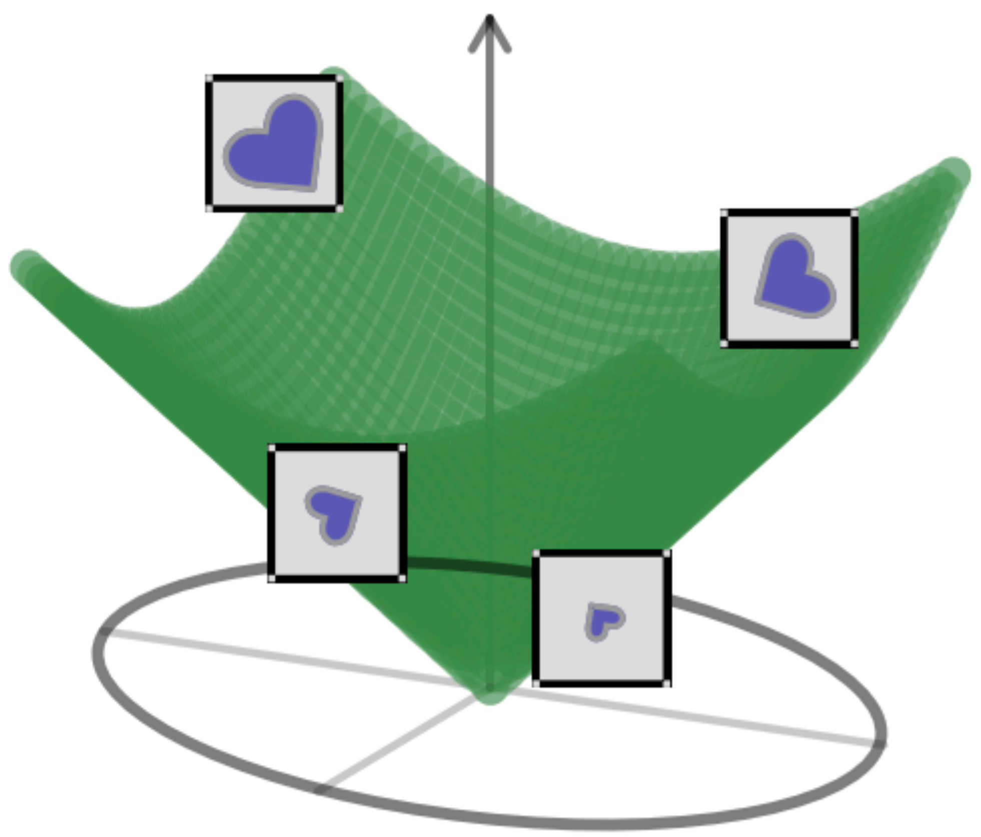
$X \sim p_{\text{model}}(\mathbf{x})$

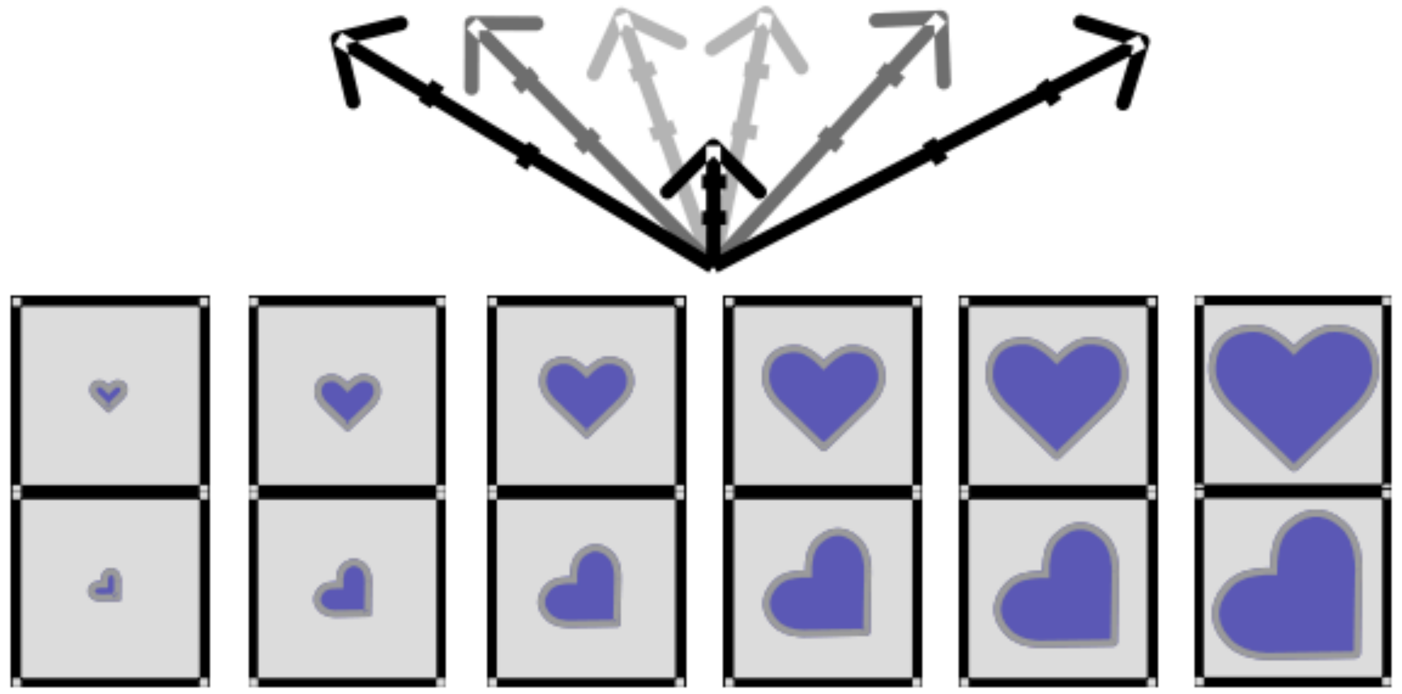
Simplices, increasing proximity

Wasserstein RLT aggregated from n runs

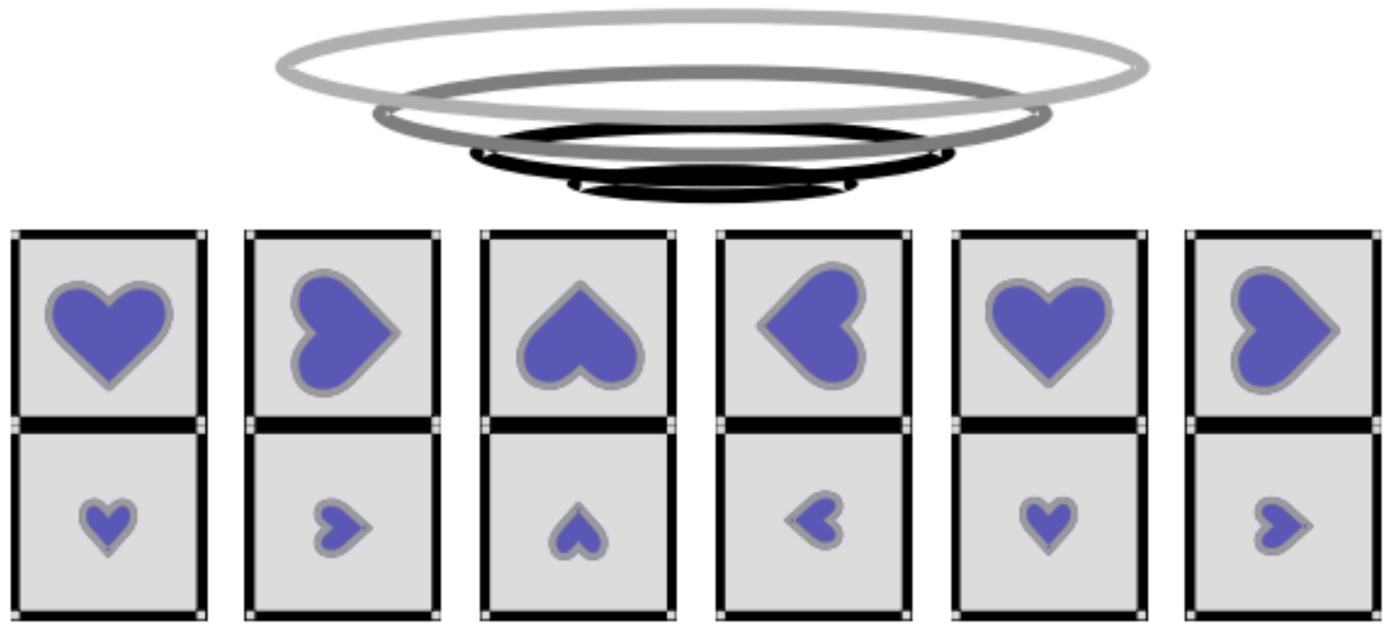
Manifold subfigure (a) is inspired by Hanson (1994)

Data points and corresponding simplices subfigures (b)-(c) by Khrulkov and Oseledets (2018)

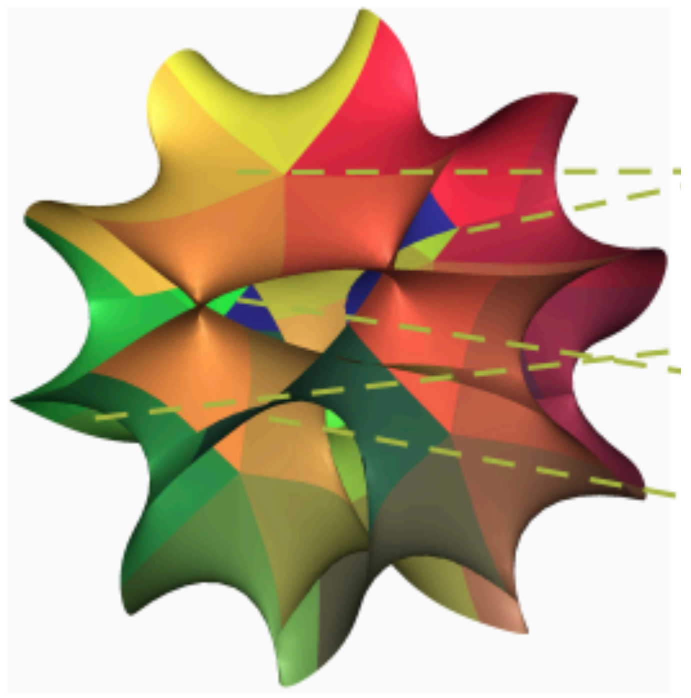




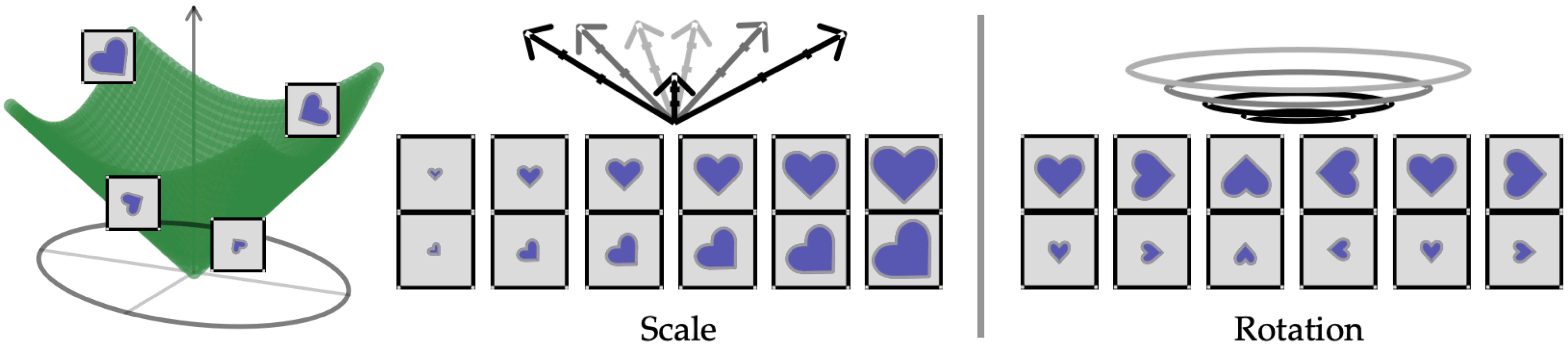
Scale

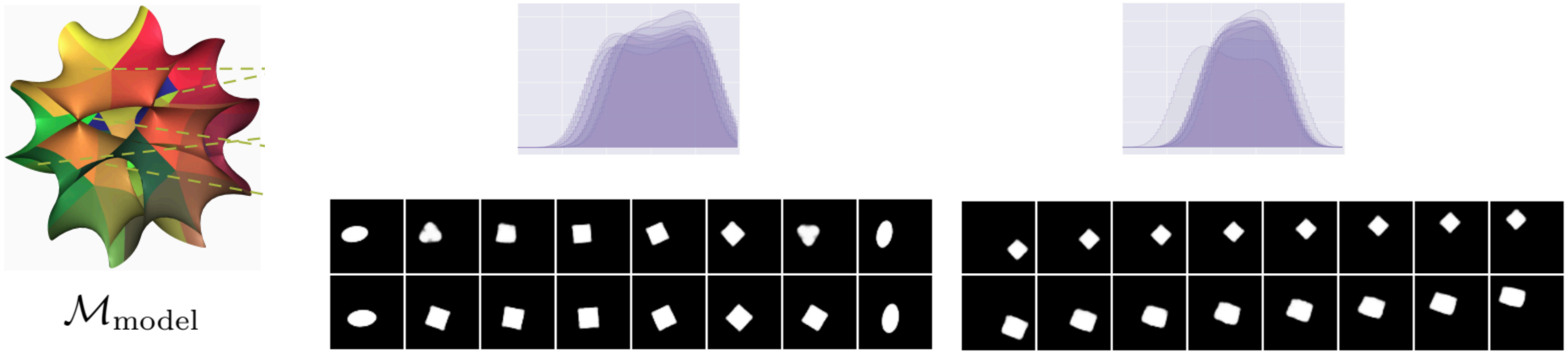


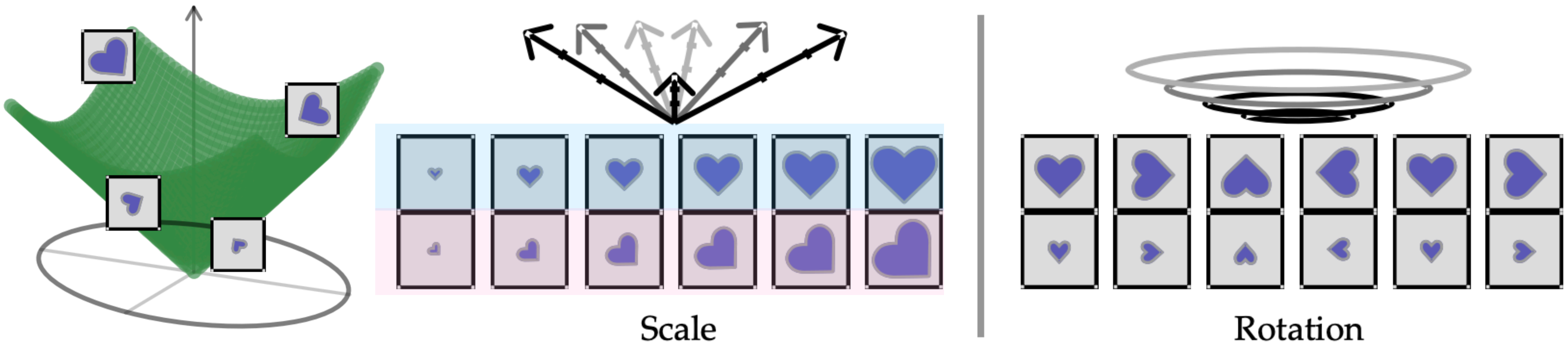
Rotation

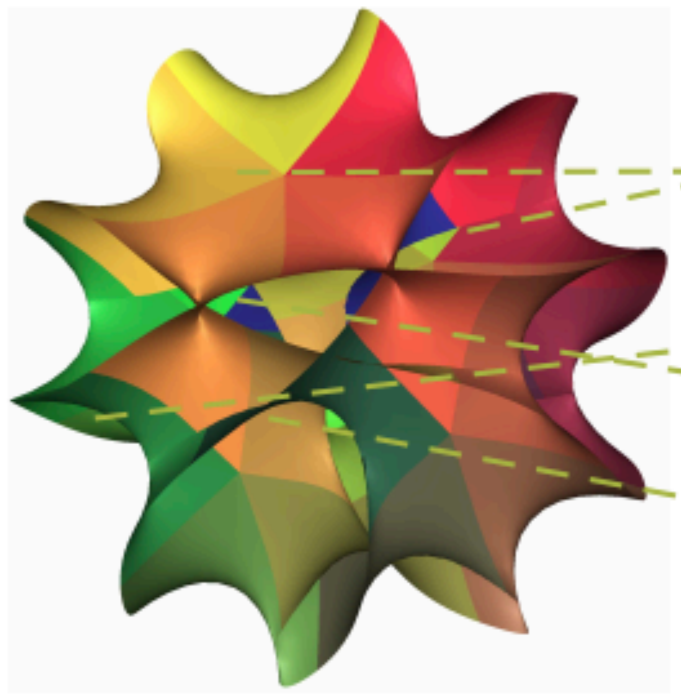


$\mathcal{M}_{\text{model}}$

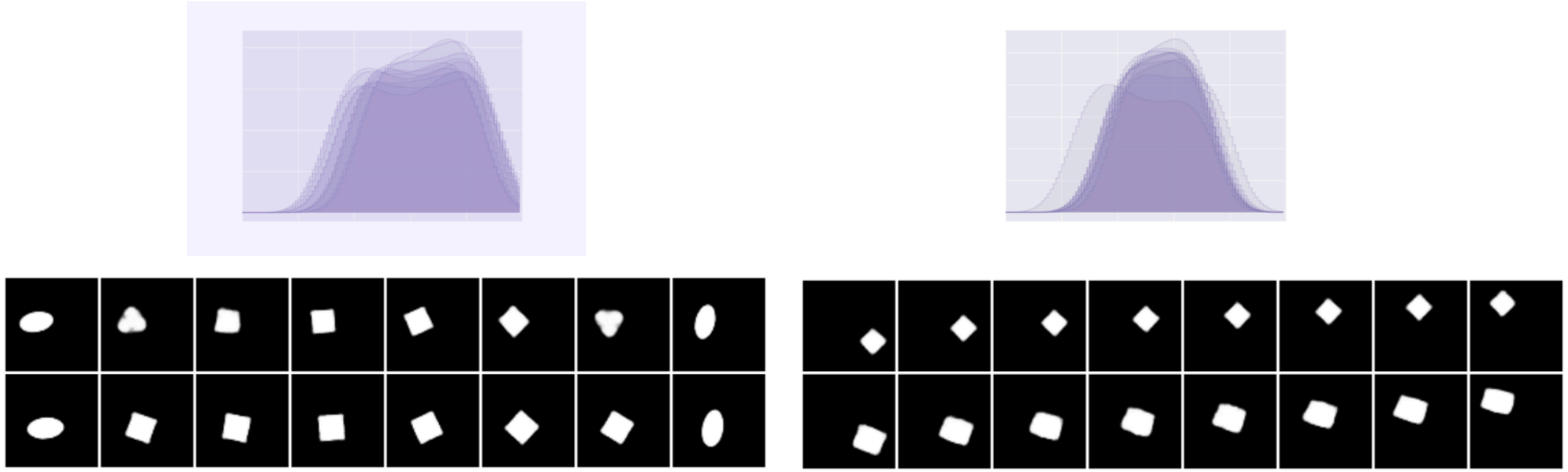

Scale
Rotation

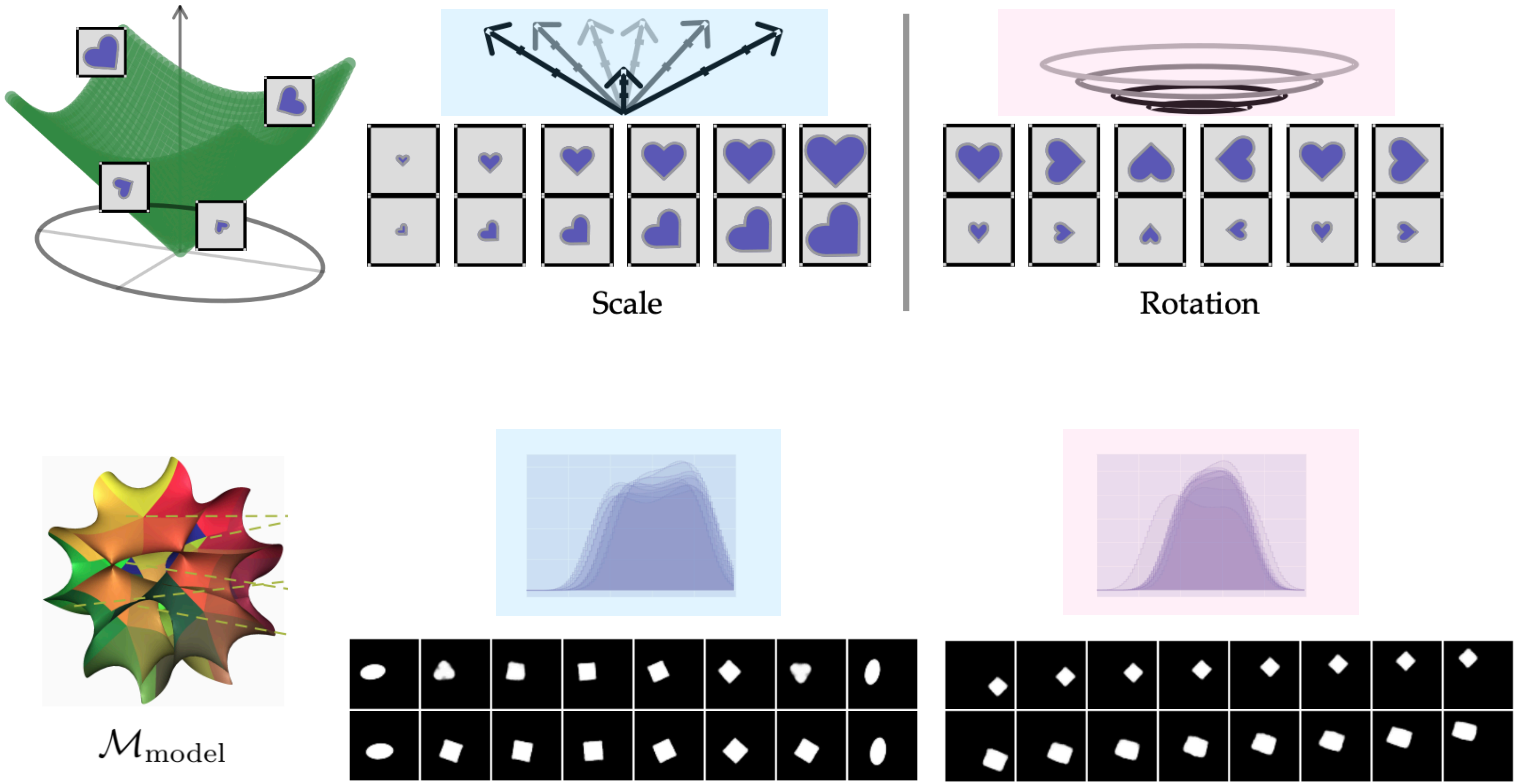

$\mathcal{M}_{\text{model}}$



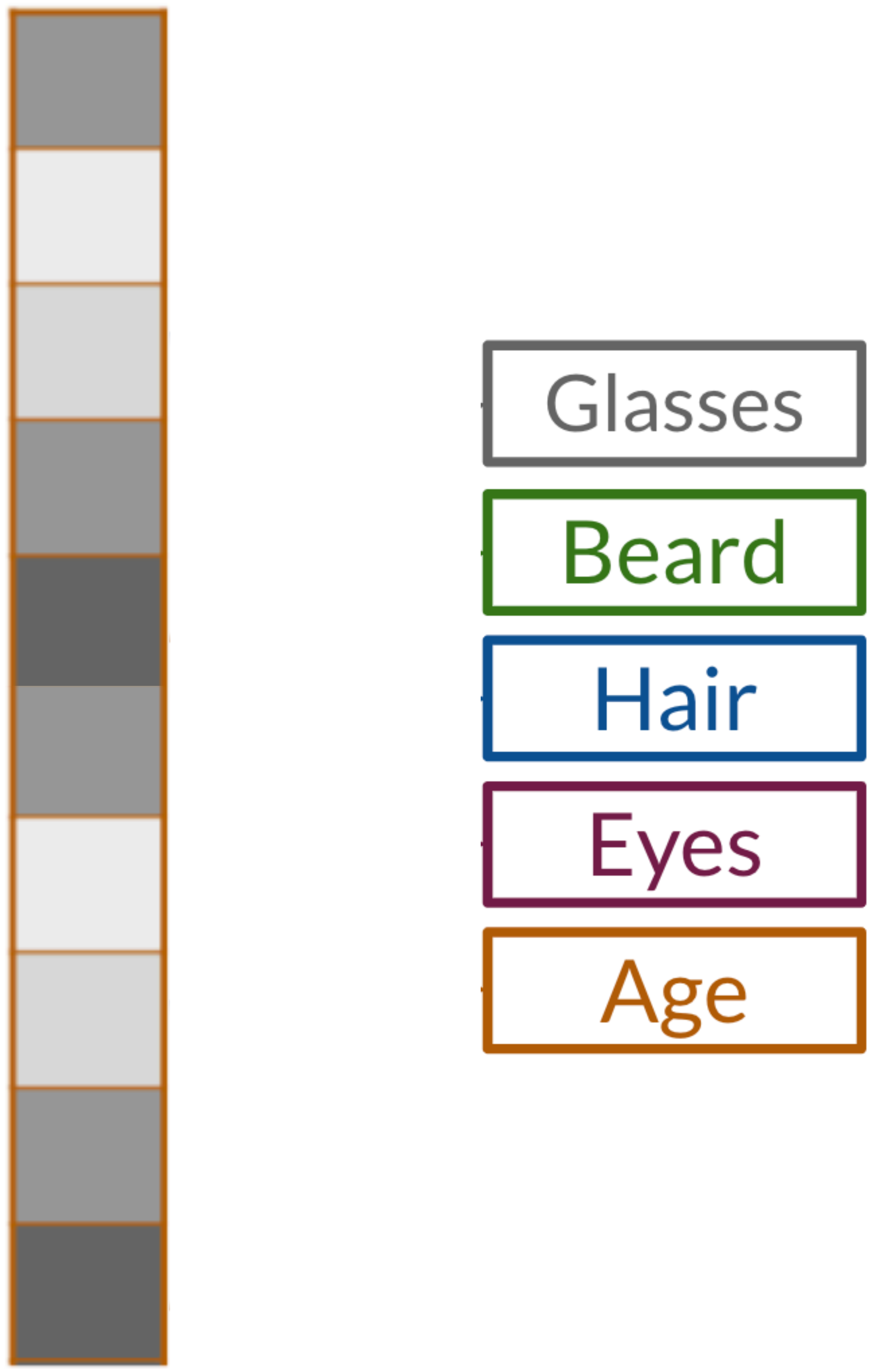


$\mathcal{M}_{\text{model}}$





Scale
Rotation
$\mathcal{M}_{\text{model}}$



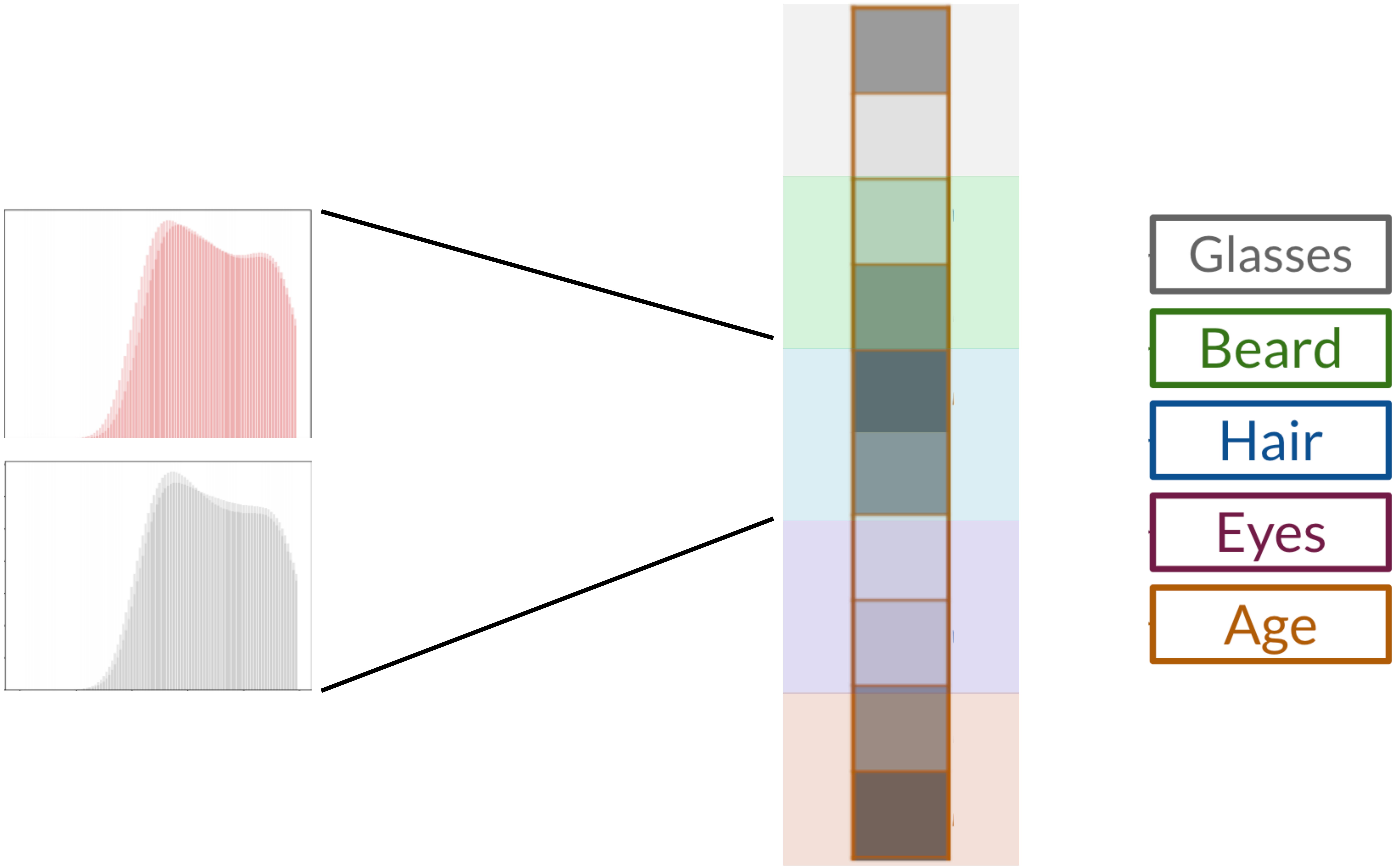
Glasses

Beard

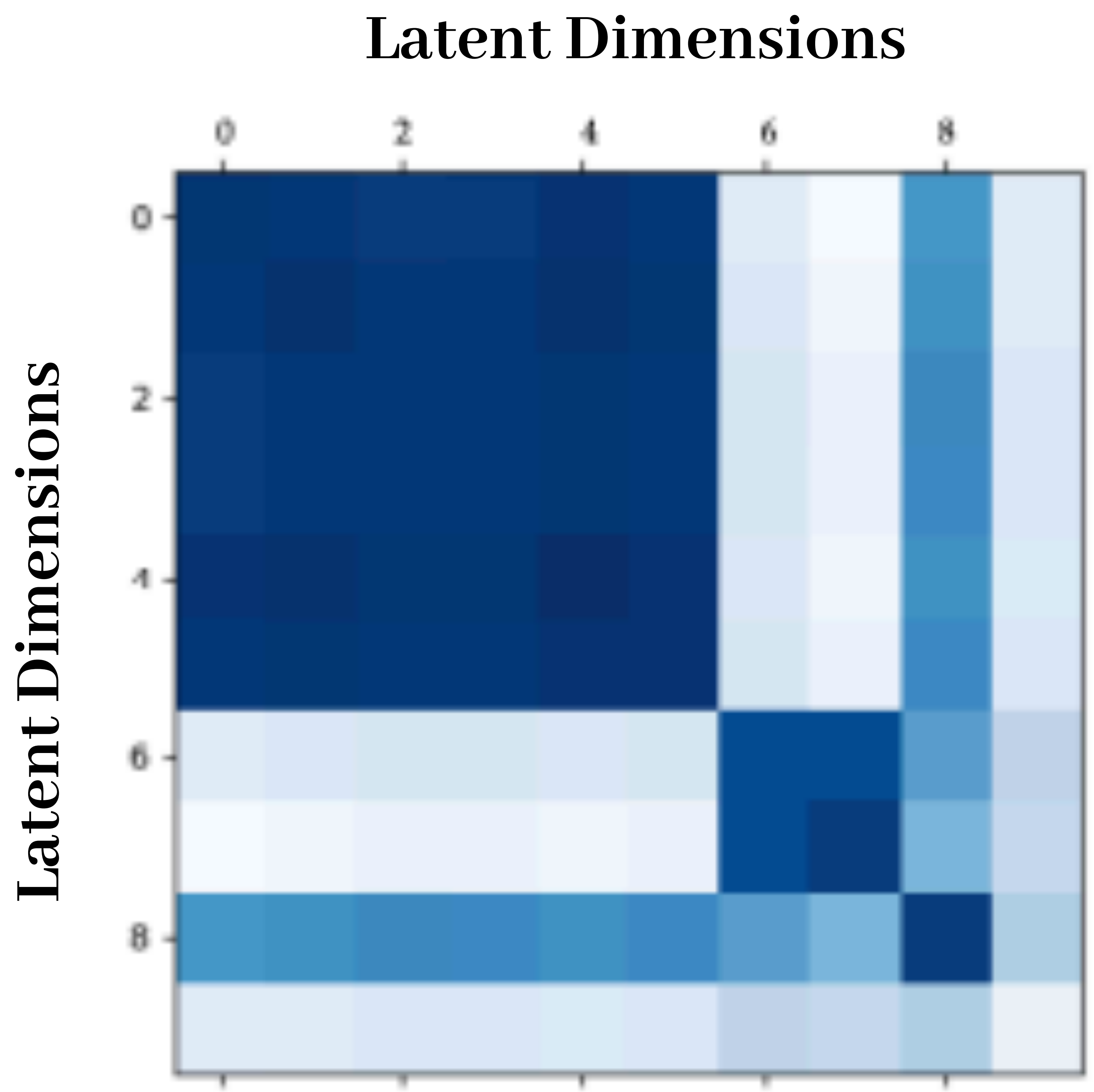
Hair

Eyes

Age


Glasses
Beard
Hair
Eyes
Age

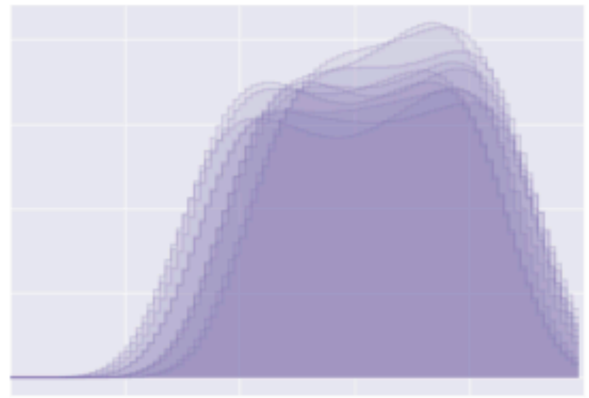
dSprites β-VAE (B)

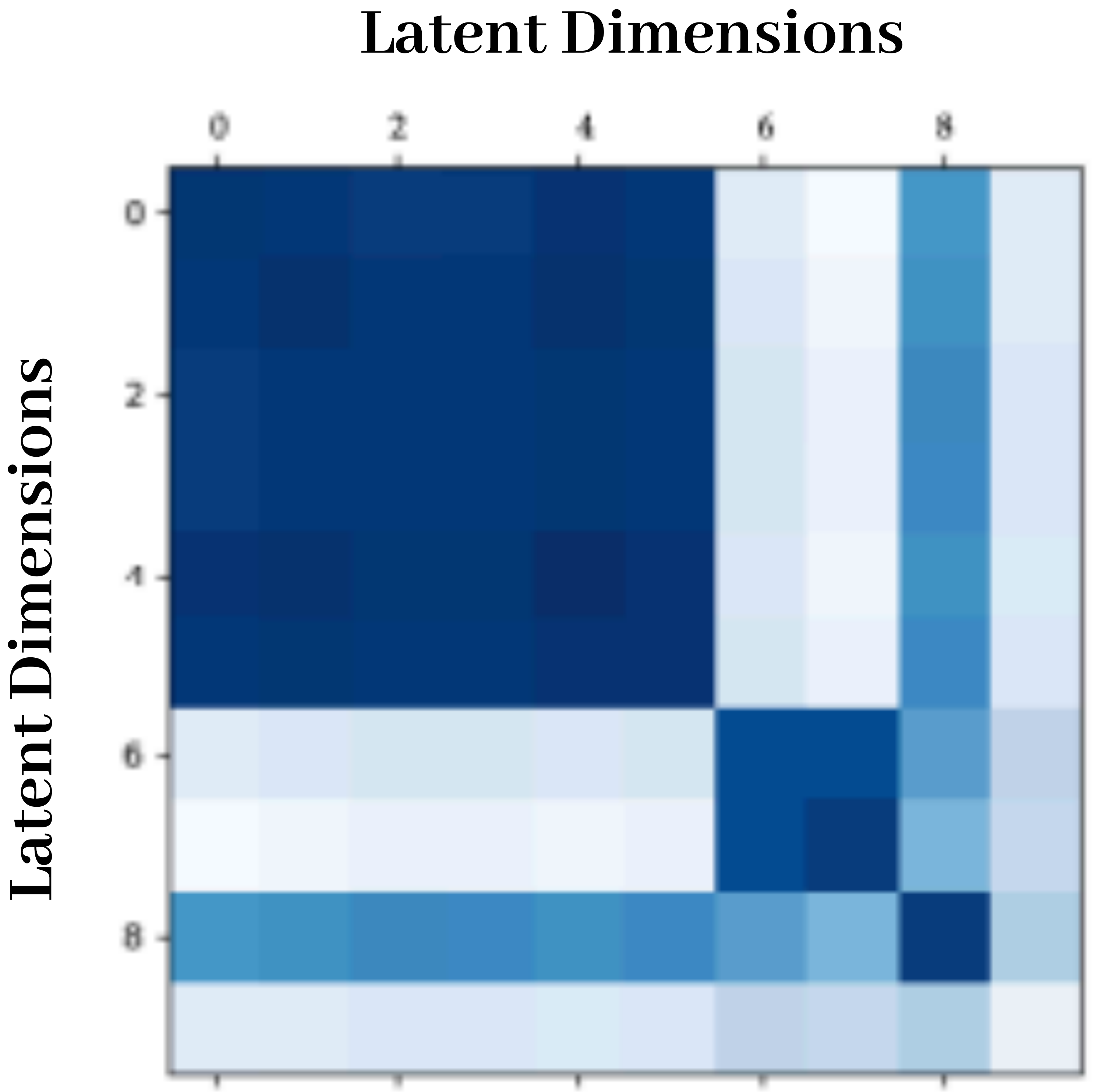


dSprites β-VAE (B)

Intracluster

Inside Diagonal Clusters

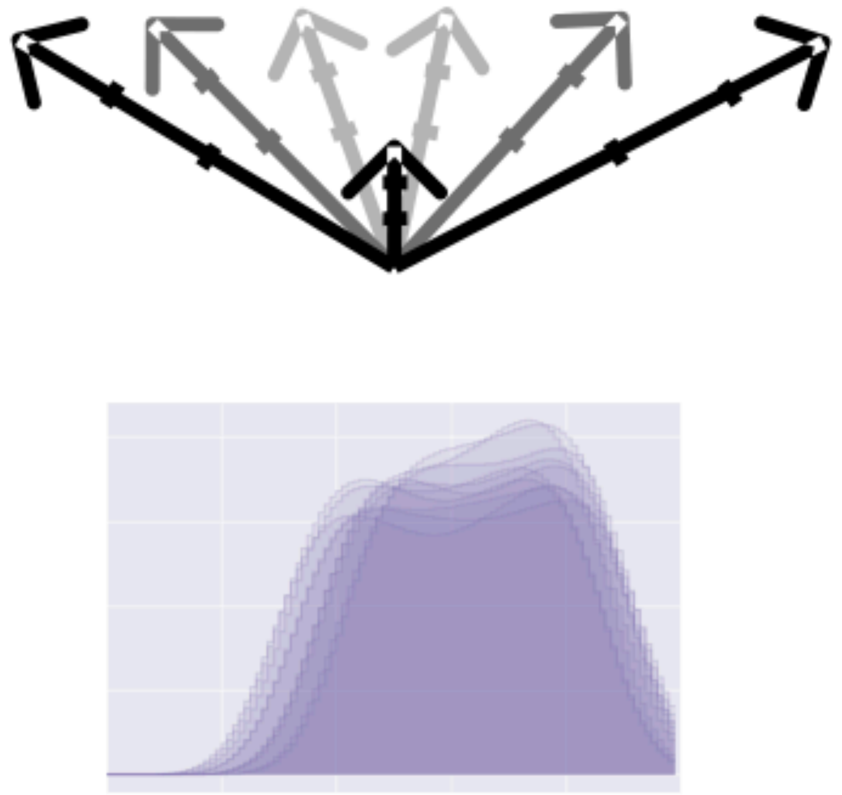


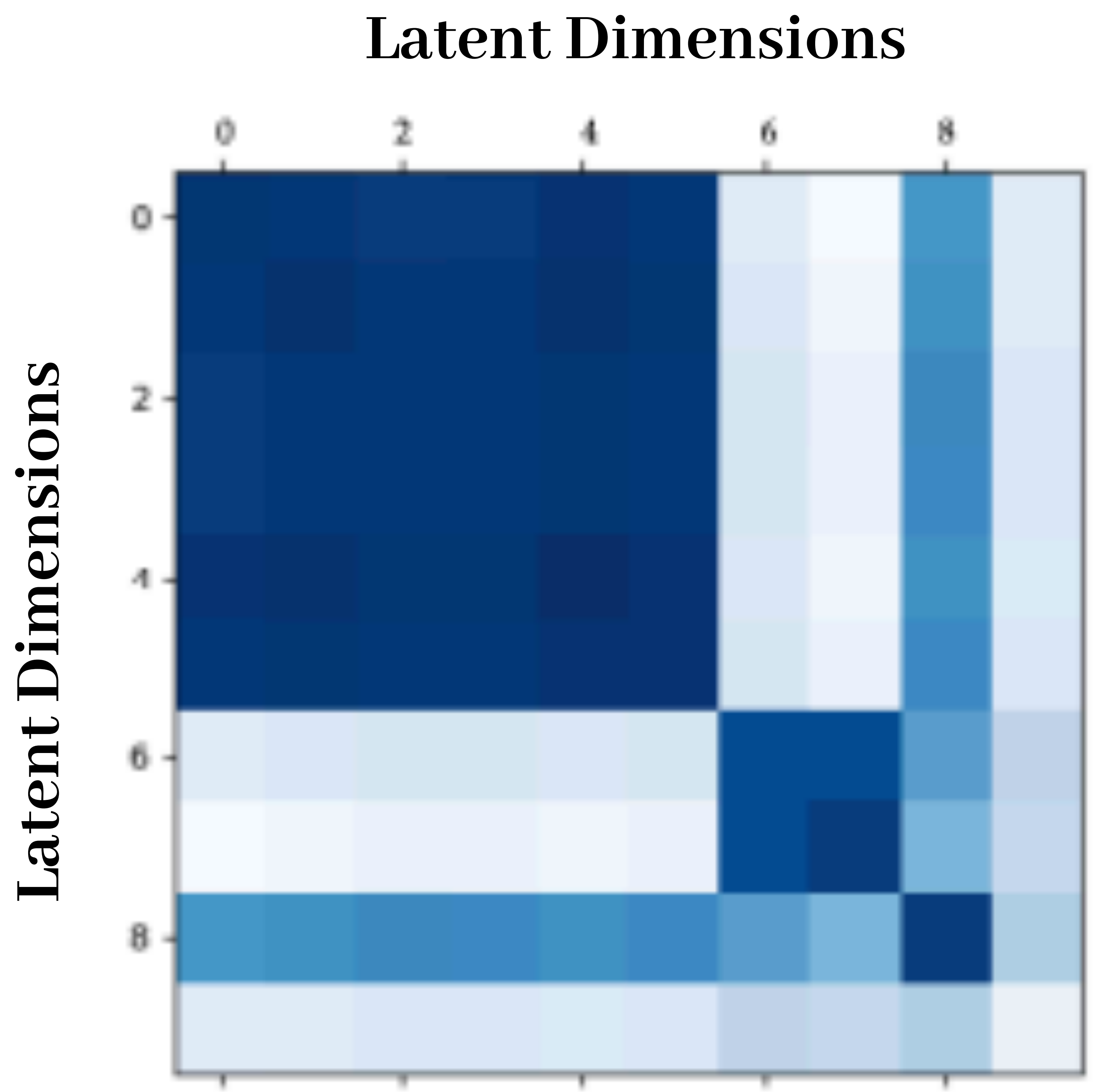


dSprites β-VAE (B)

<u>Intracluster</u>

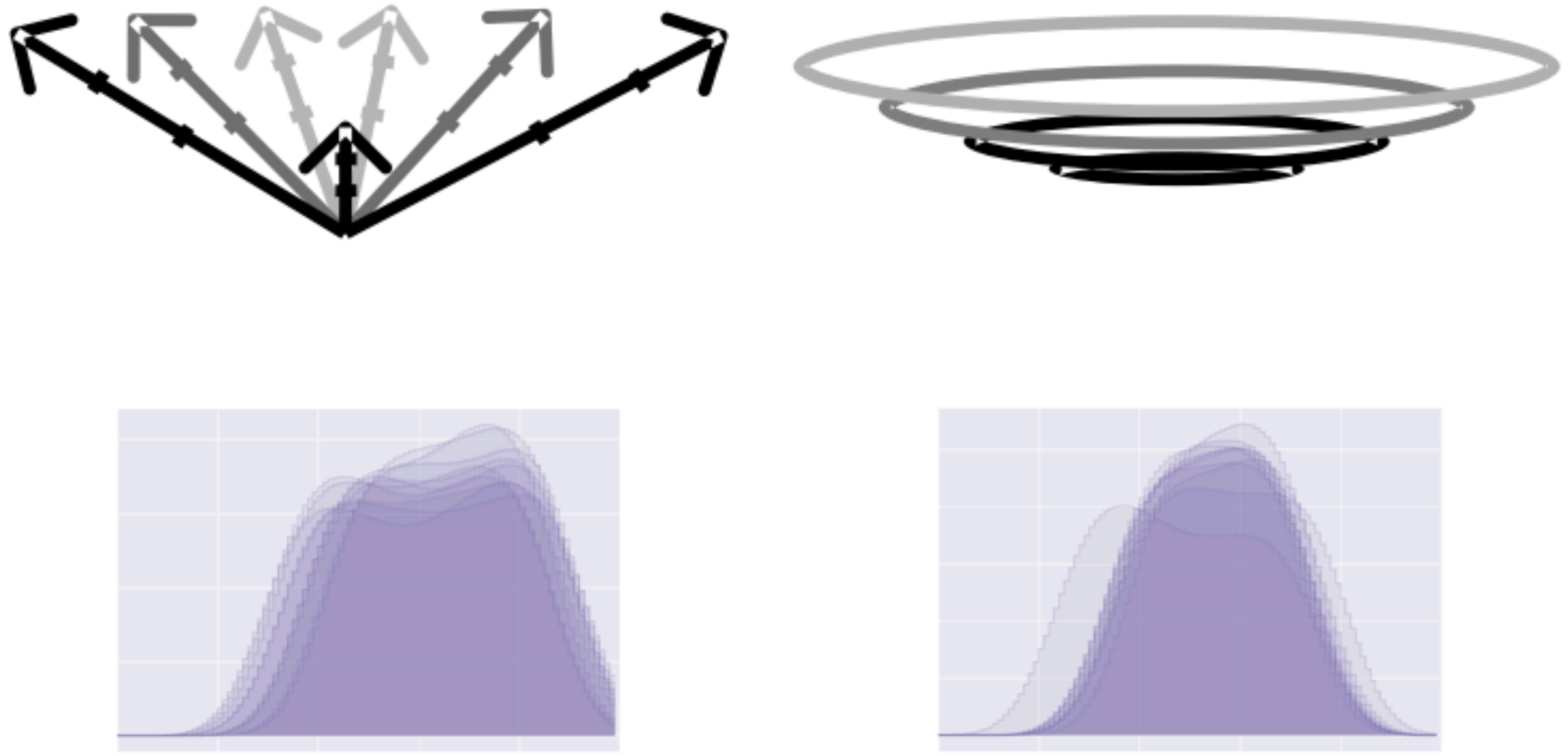
Inside Diagonal Clusters



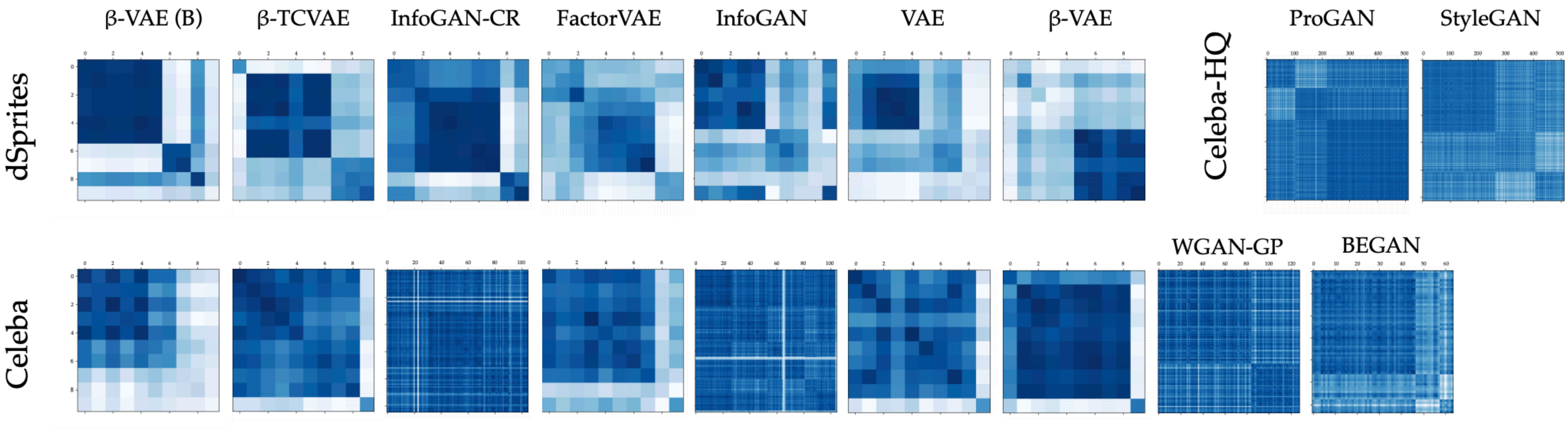


<u>Extracluster</u>

Outside Diagonal Clusters



Spectrally Coclustered Topological Similarity Matrices



Prior methods

1. Depends on the **architecture** specifically with an **external model**, e.g. encoder and/or classifier
2. **Supervision** is required for a classifier
3. Tuned to a **specific dataset**, e.g. custom preprocessing on face images

Ours

1. Uses an **intrinsic property** of a generative model, without reliance on external models or custom architectures
2. **Unsupervised** and supervised variants both available
3. Procedure can be **applied across datasets** — and architectures, as above