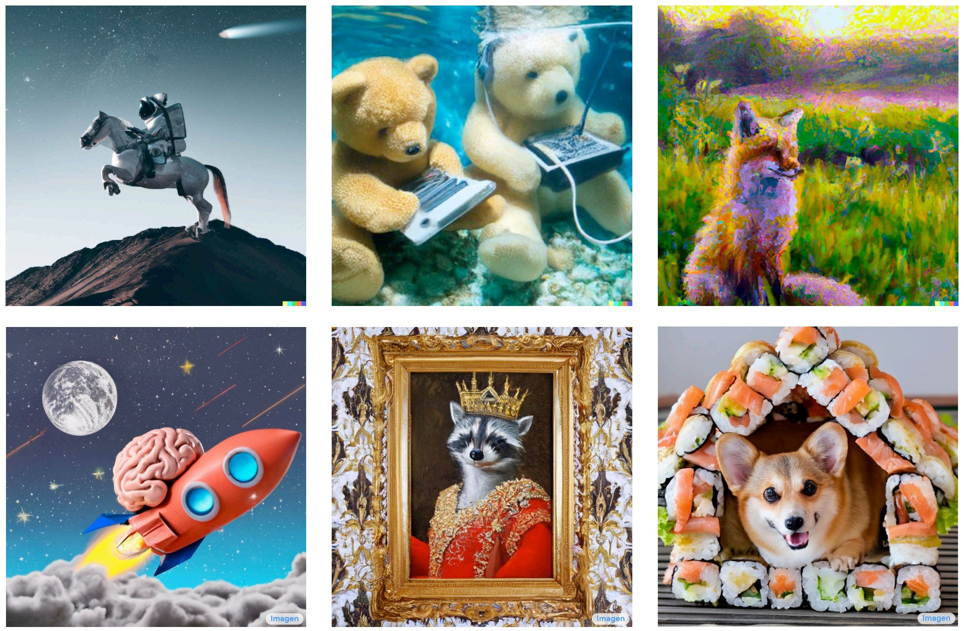
Emergence and Evolution of Interpretable Concepts in Diffusion Models

Berk Tinaz* **Zalan Fabian*** **Mahdi Soltanolkotabi**
Dept. of Electrical and Computer Engineering
University of Southern California
Los Angeles, CA, USA
`tinaz@usc.edu` `fabian.zalan@gmail.com` `soltanol@usc.edu`

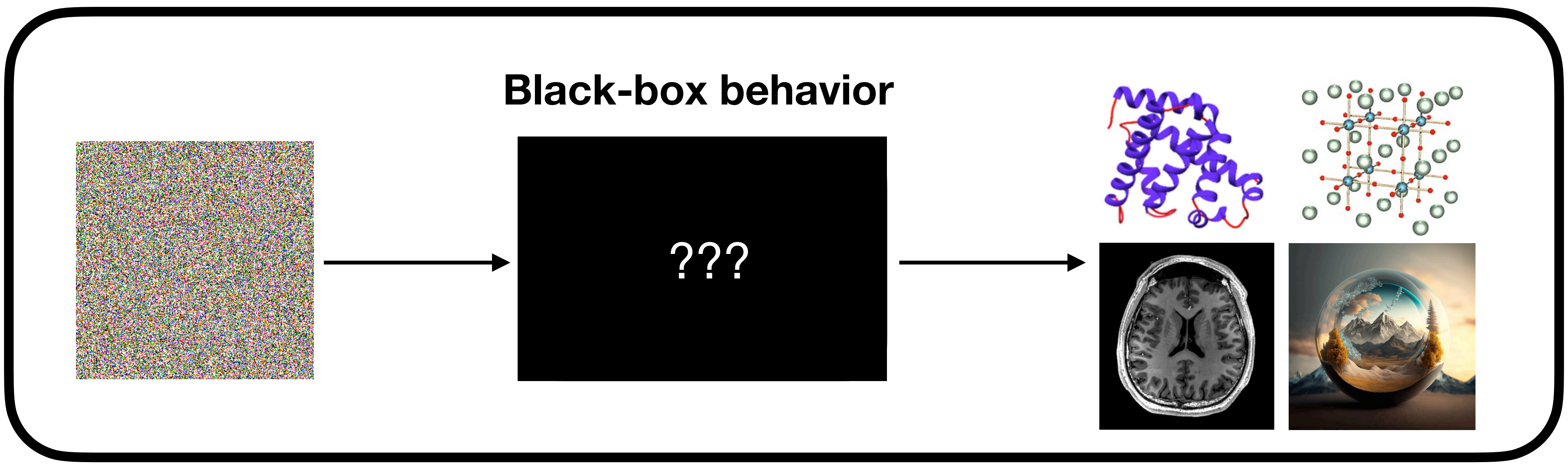
USC Center on AI Foundations for the Sciences

NEURAL INFORMATION PROCESSING SYSTEMS

Interpretable Diffusion



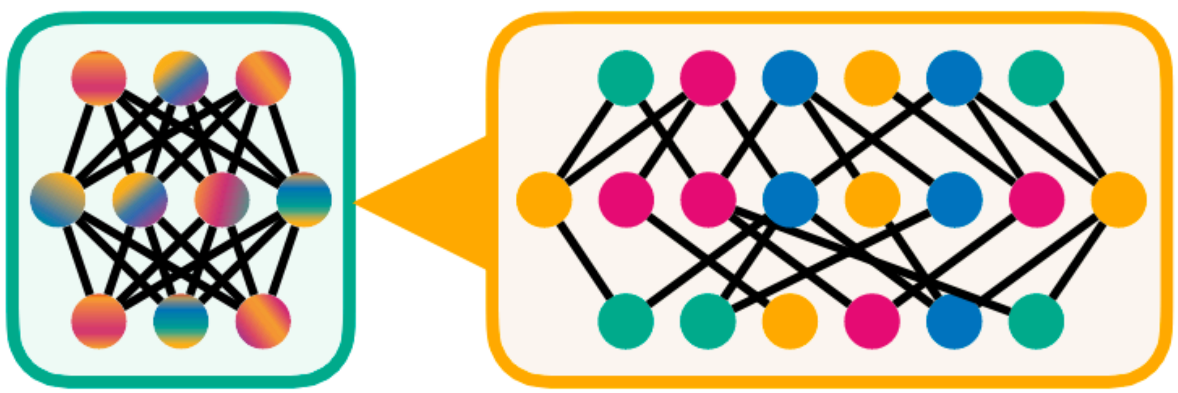
Diffusion models are the **powerhouse** behind modern generative AI.



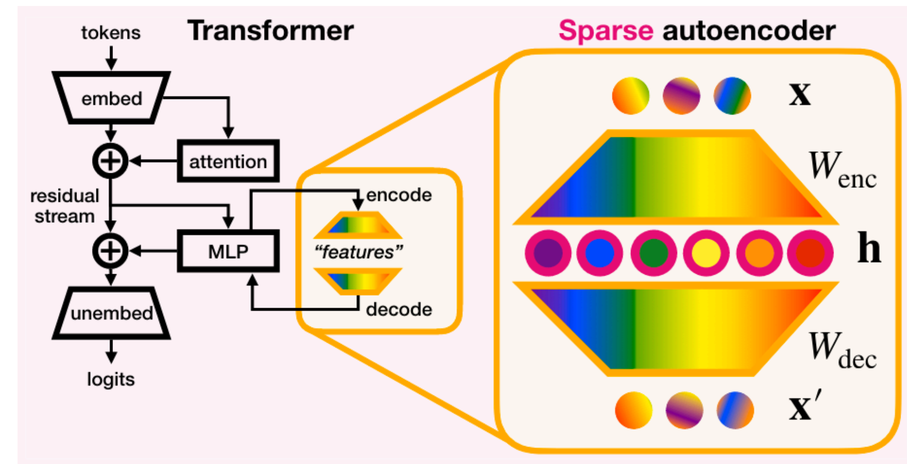
Our understanding of the inner mechanisms of diffusion models remains **limited**.

Sparse Autoencoders

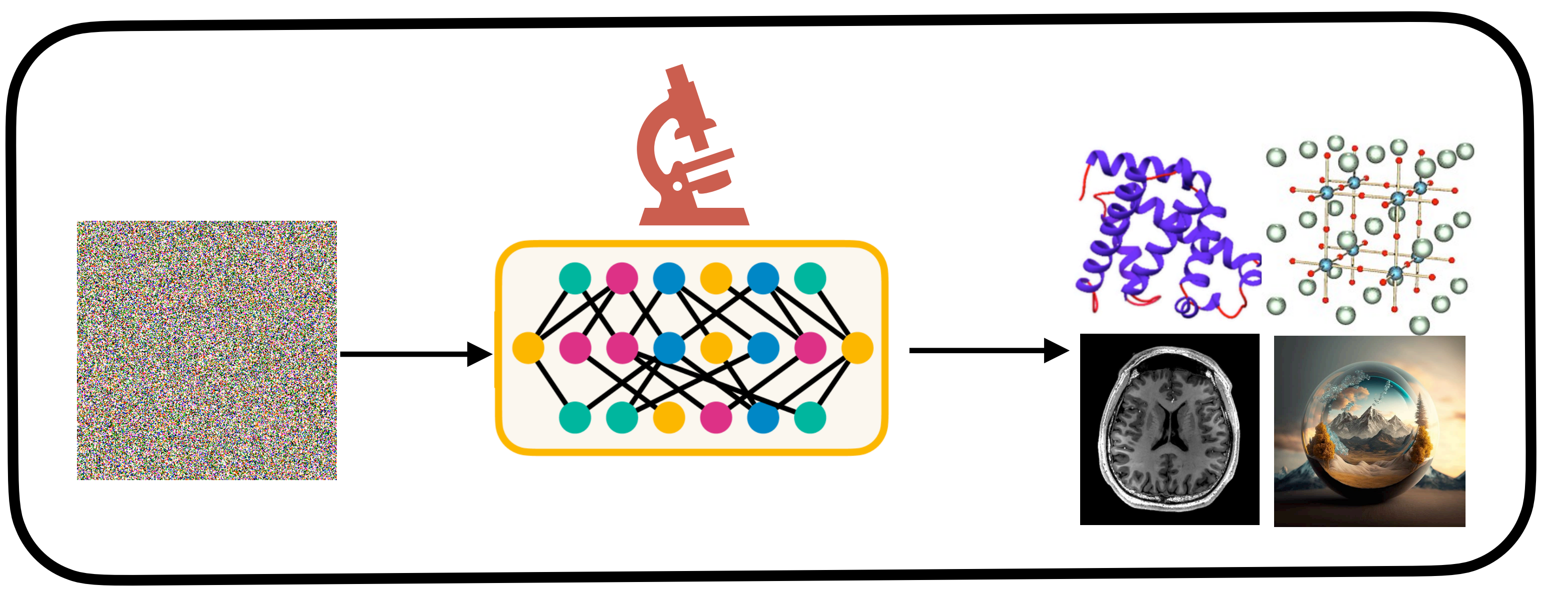
Mechanistic interpretability through Sparse Autoencoders (SAEs)



Observed model **Hypothetical disentangled model**



Diffusion models under the lens of SAEs



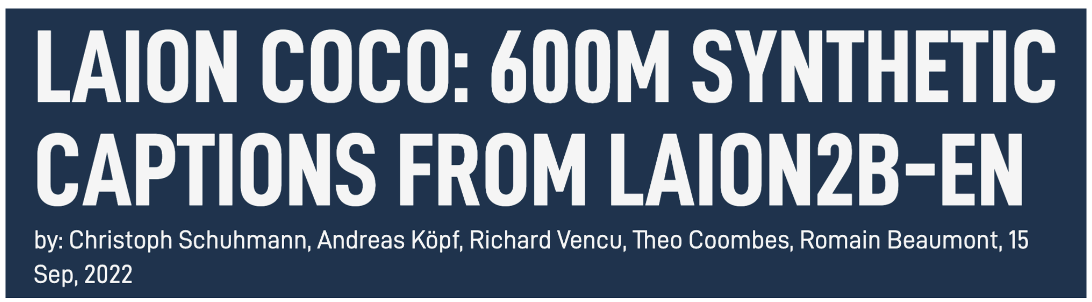
Goals:

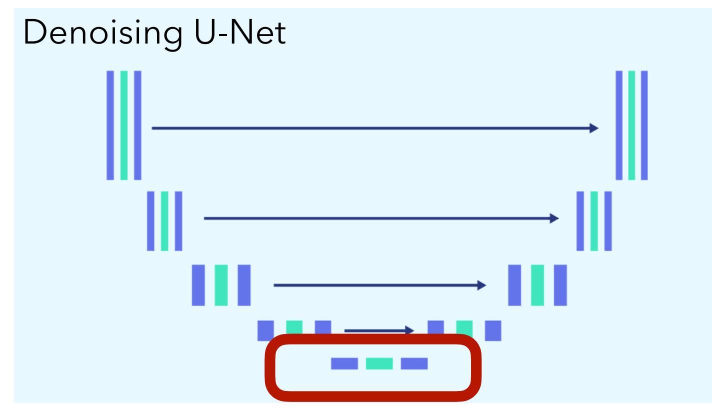
1. Understand the features diffusion models learn
2. Identify causal mechanisms
3. Achieve fine-grained control of the generative process

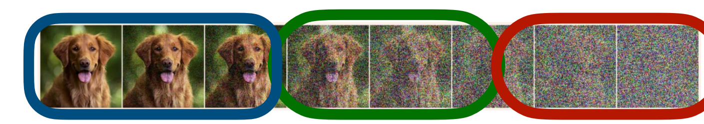
SAE figures from Bereska et al., Mechanistic Interpretability for AI Safety -- A Review, arXiv:2404.14082

Training Sparse Auto-encoders

Use diverse text prompts and collect activations from the bottleneck layer of <u>**Stable Diffusion v1.4,**</u> at different time steps of the reverse diffusion.







<u>Training:</u>

Pick top K values, set rest to 0.

Encoder: $$\boldsymbol{z} = \mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathrm{TopK}\left(\mathrm{ReLU}\left(\boldsymbol{W}_{enc}(\boldsymbol{x} - \boldsymbol{b})\right)\right)$$

Decoder: $$\hat{\boldsymbol{x}} = \mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{z}) = \boldsymbol{W}_{dec}\boldsymbol{z} + \boldsymbol{b}$$

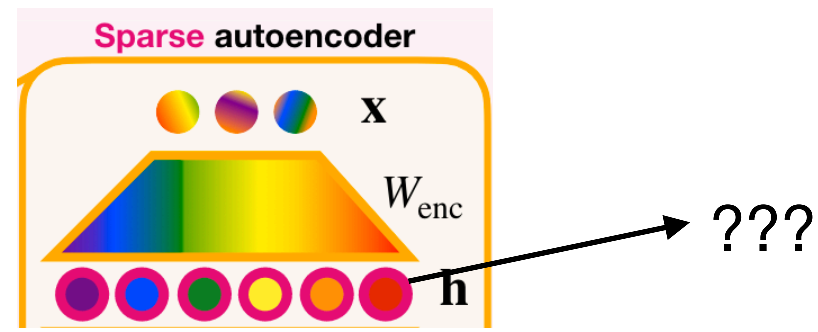
$$\mathcal{L}_{rec}(\boldsymbol{\theta}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 + \boxed{\alpha\mathcal{L}_{aux}(\boldsymbol{\theta})}$$

Prevent “dead” neurons

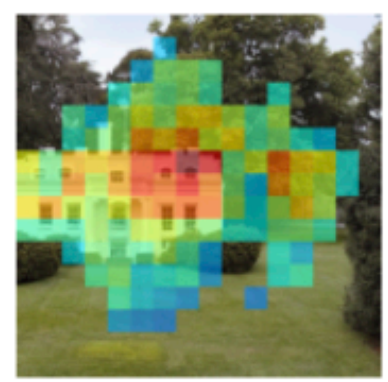
Gao, Leo, et al. "Scaling and evaluating sparse autoencoders." *arXiv preprint arXiv:2406.04093* (2024).

Labeling SAE Features

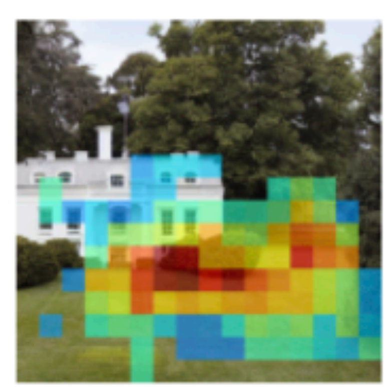
How to automatically label SAE features?



Assign list of objects to each SAE feature using an open-source image segmentation pipeline!



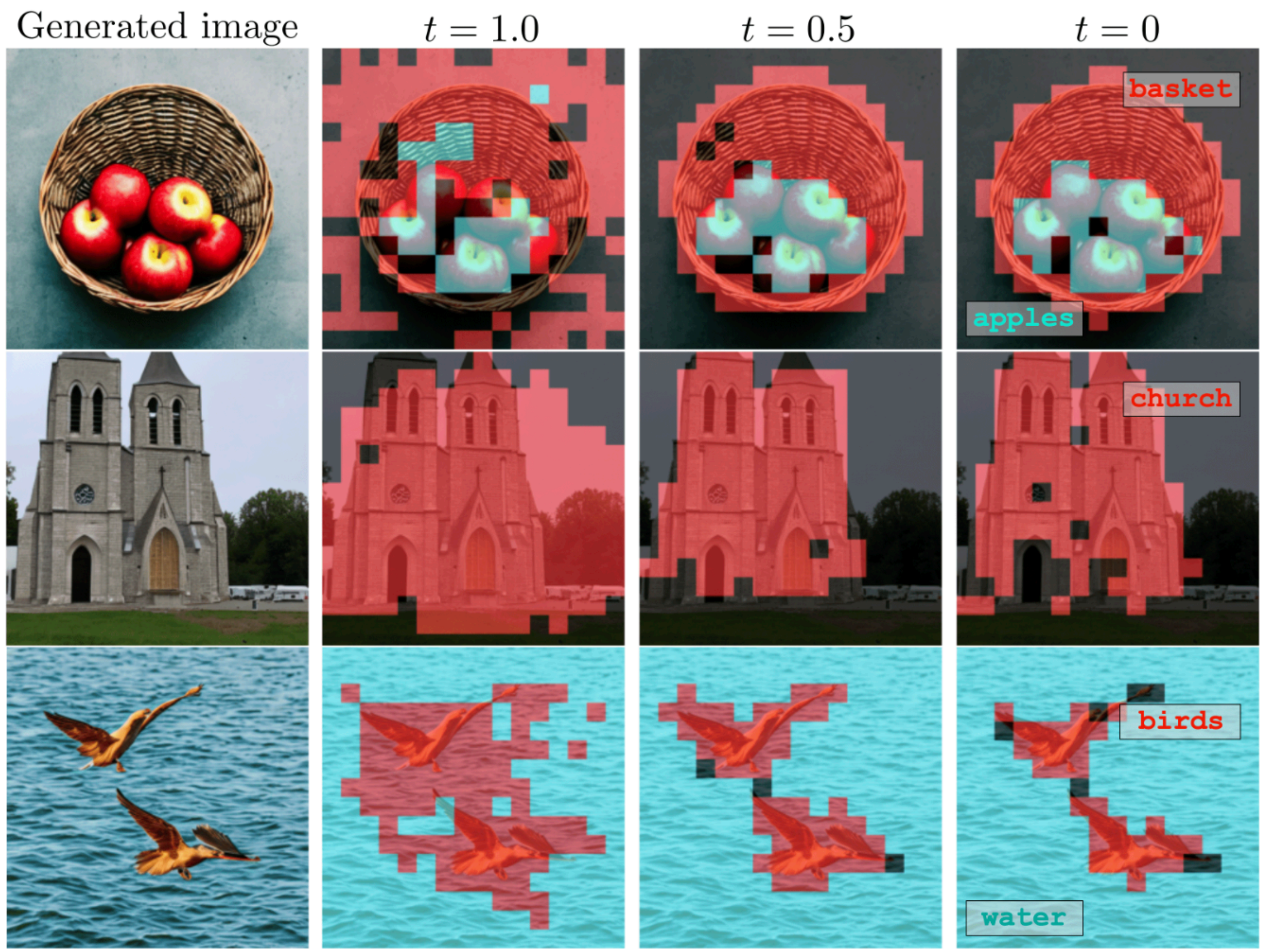
CID: 4657
white house, building



CID: 2577
grass

Semantic Segmentation

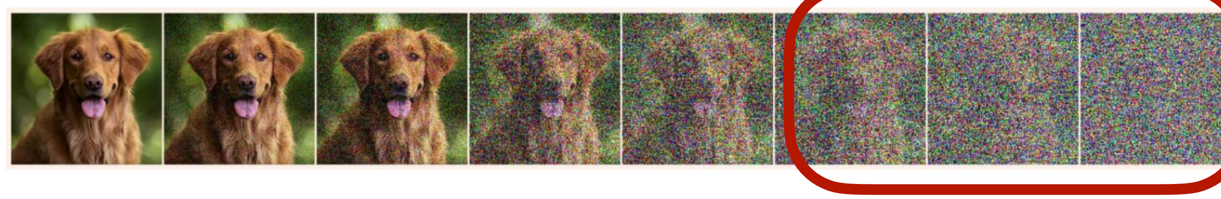
When does the image layout emerge during sampling?

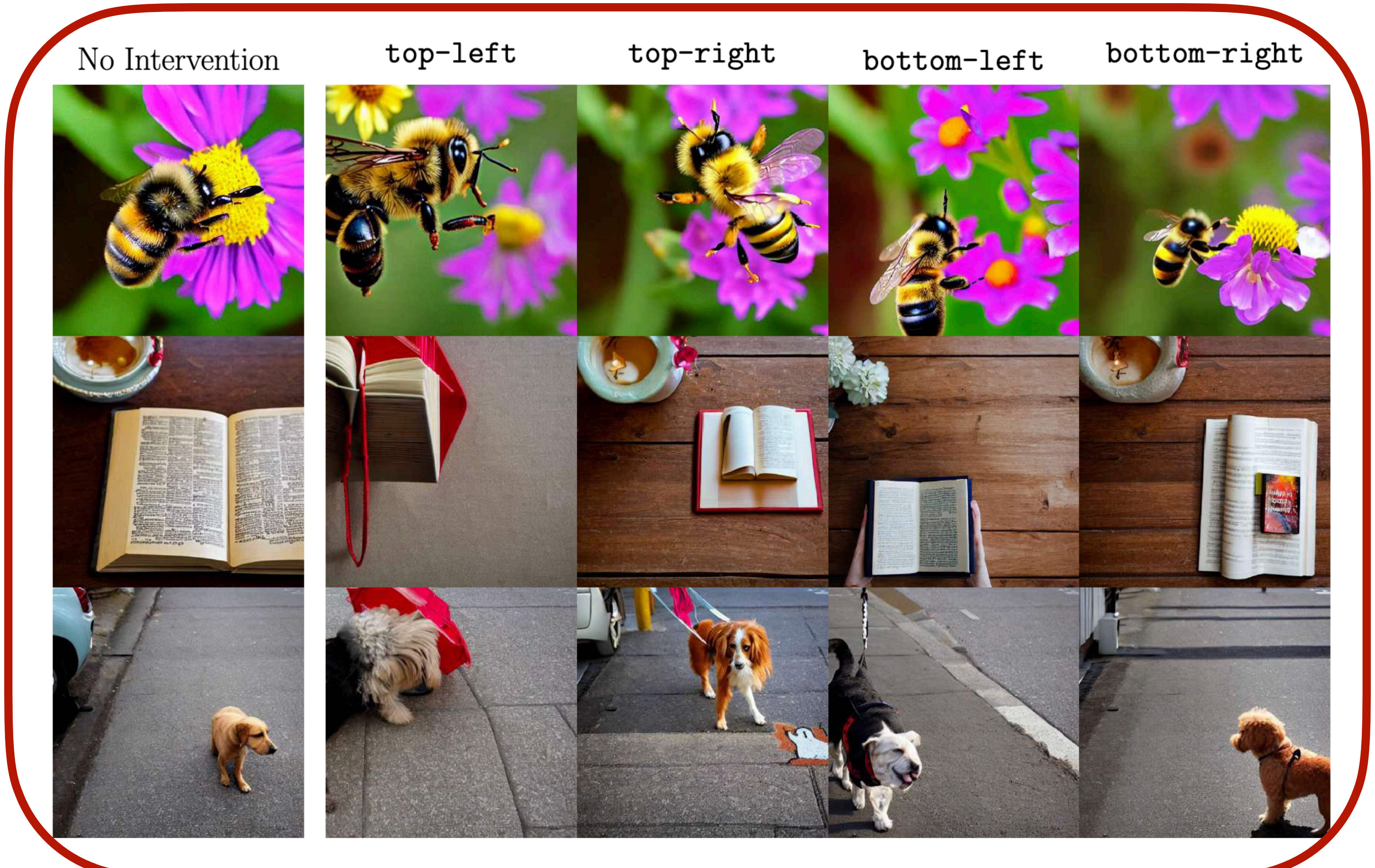


We can **predict** the image layout even **before** the first diffusion step is completed and prediction **gets better** as time progresses.

Spatially Targeted Interventions

Interventions at

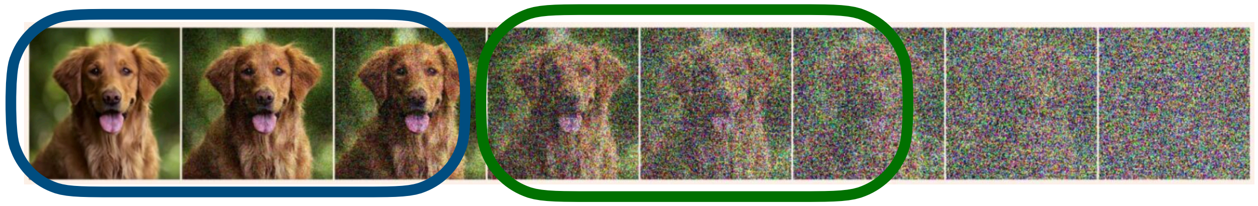


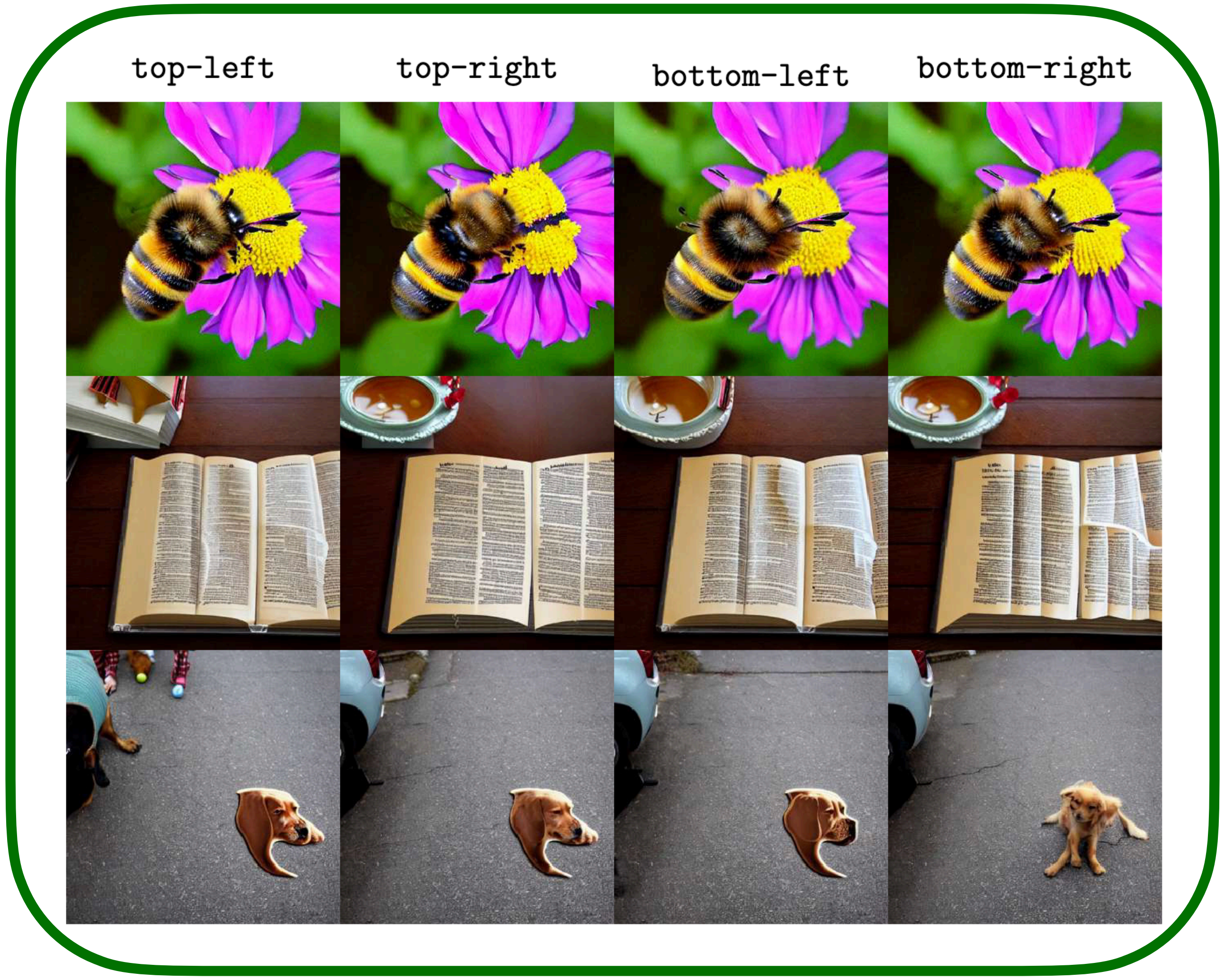


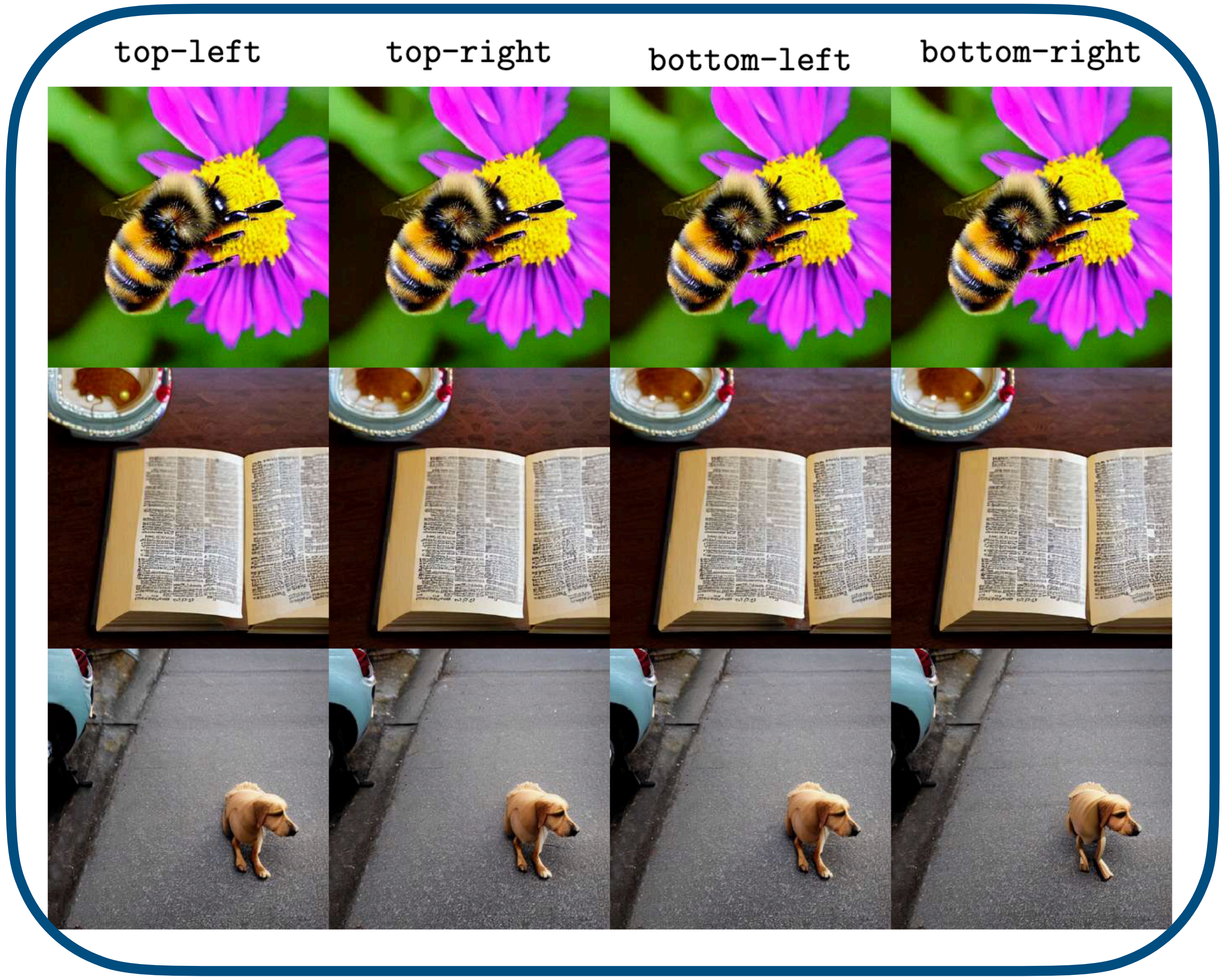
We can **restrict** objects to the specified quadrant of the image when intervened in **early stages**.

Spatially Targeted Interventions

Interventions at







Our interventions are **unsuccessful** in the **middle** and **final** stages of reverse diffusion.

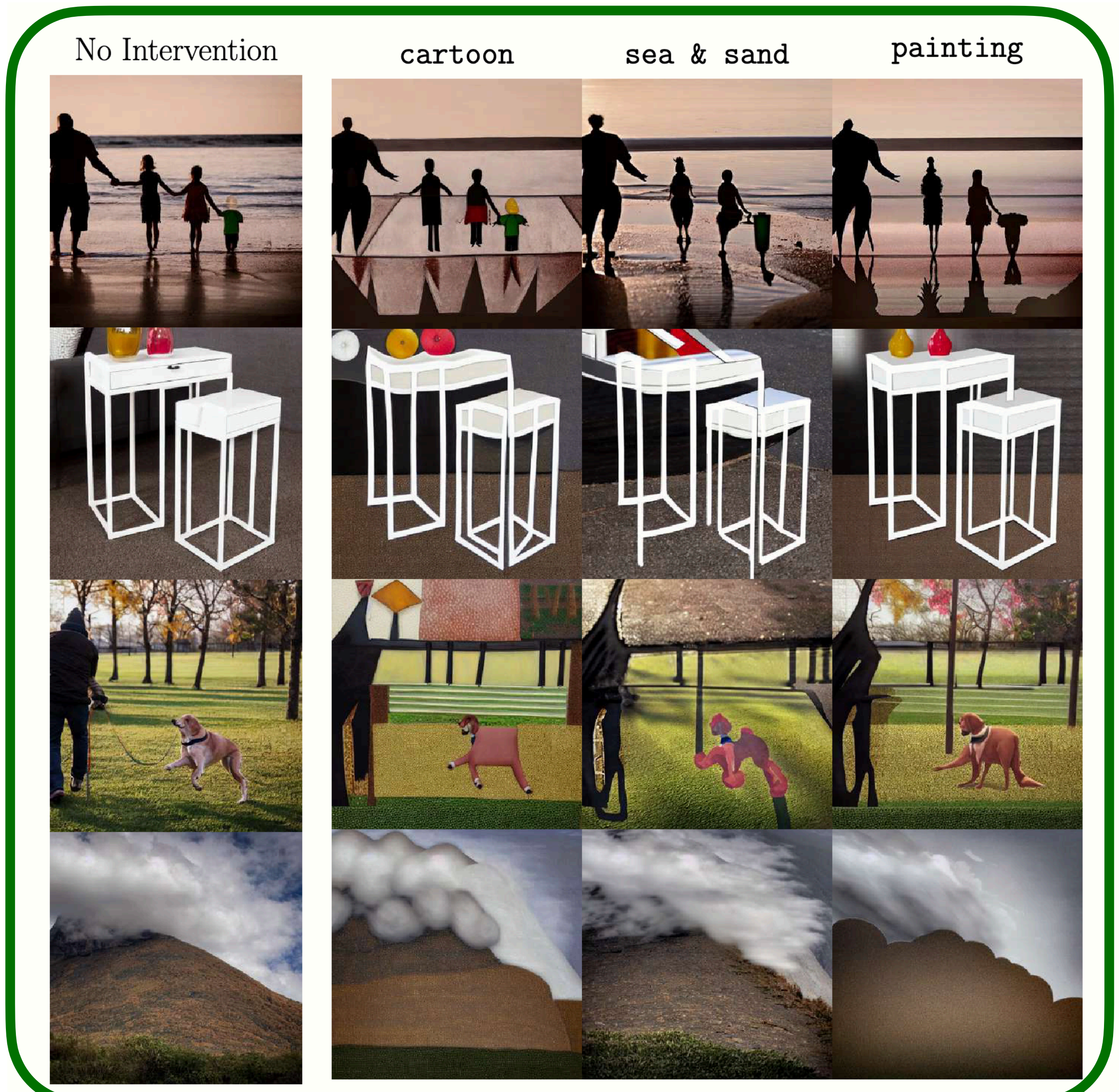
Global Interventions (Early Stage)

Interventions at

Edits in **initial steps** of reverse diffusion drastically modify the **broad composition** of the image.

Global Interventions (Middle Stage)

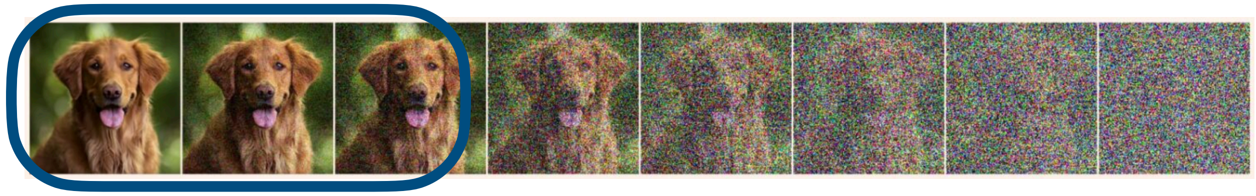
Interventions at

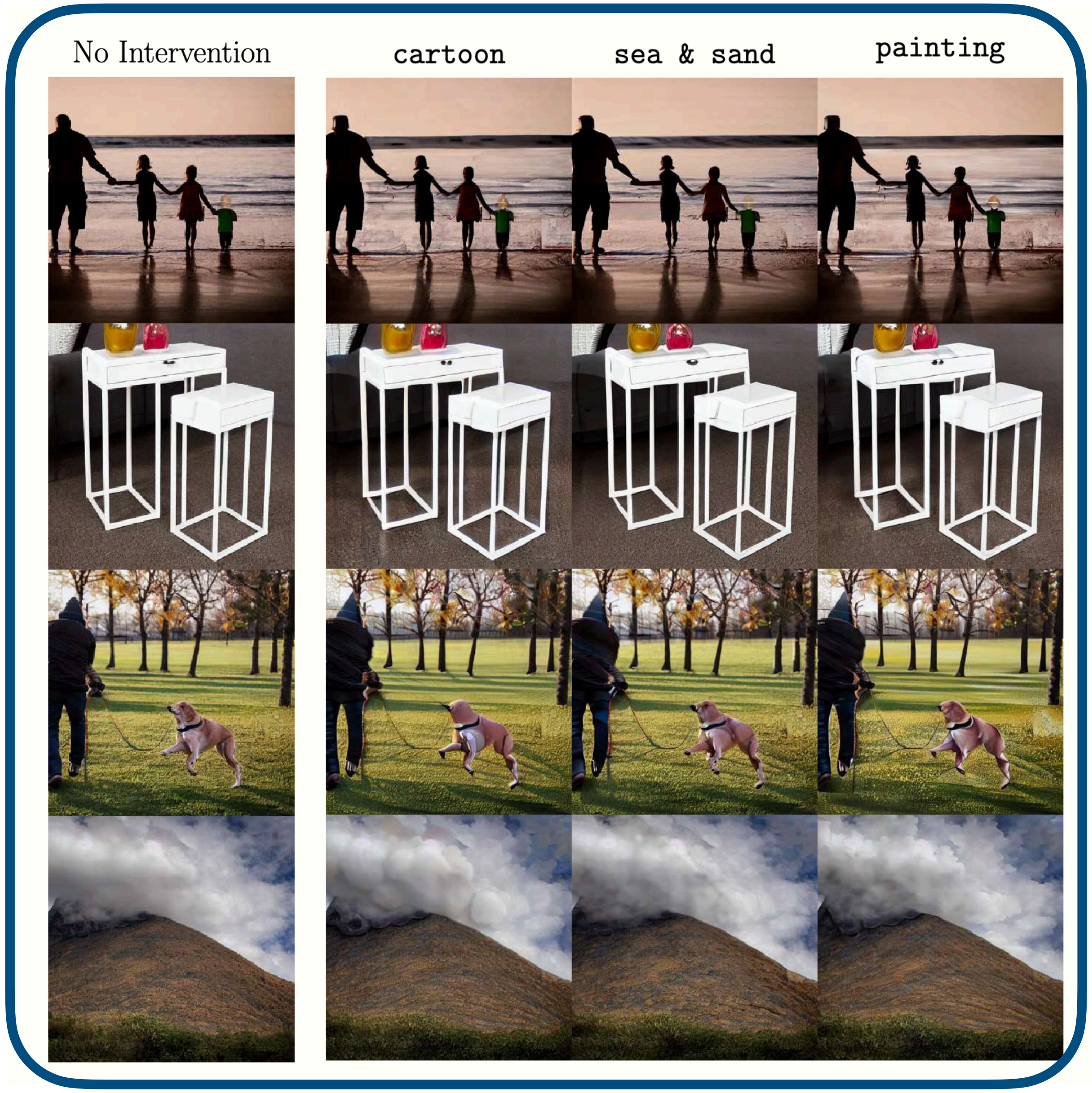


Middle-stage interventions successfully manipulate image **style** without interfering with **image composition.**

Global Interventions (Final Stage)

Interventions at





Global interventions in the **final** stages have **no effect** on **<u>style</u>** or **<u>composition</u>**, results <u>only</u> in minor textural changes.

Thank you for your attention!

GitHub

Paper

https://github.com/berktinaz/stable-concepts

Contact: tinaz@usc.edu