Finding NeMo: Localizing Neurons Responsible For Memorization in Diffusion Models

Dominik Hintersdorf *, [1, 2] Lukas Struppek *, [1, 2] Kristian Kersting [1, 2, 3, 4] Adam Dziedzic [5] Franziska Boenisch [5]

[1] TU Darmstadt [2] German Research Center for AI (DFKI) [3] Centre for Cognitive Science [4] Hessian Center for AI [5] CISPA Helmholtz Center for Information Security

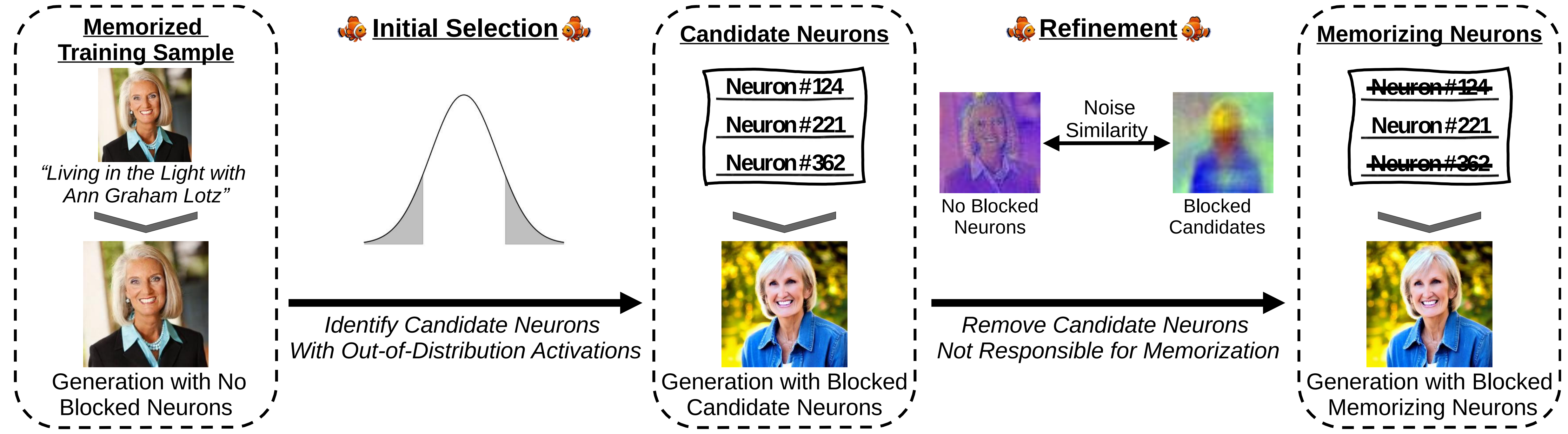
* Equal contribution

At a Glance

- **Neuron Memorization (NeMo)** localizes the memorization of training samples in diffusion models down to **individual neurons**.
- **Single neurons** within Stable Diffusion are responsible for memorizing multiple training samples.
- All memorization is confined to neurons in the **cross-attention value layers** of the U-Net's down-blocks.
- Deactivating memorization neurons **mitigates memorization and increases output diversity** without compromising image quality.

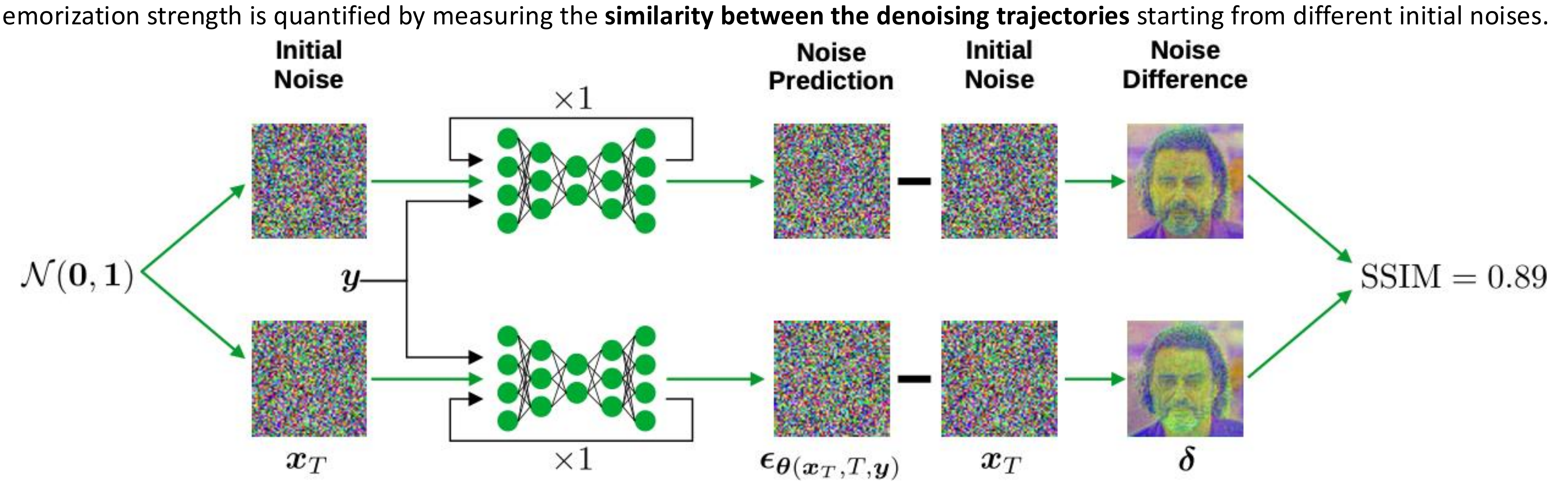
Localizing Memorization

NeMo detects candidate memorization neurons based on their **activation patterns**. The number of initially found neurons is then reduced in a second **refinement step**. This design makes NeMo **very fast and efficient** since no gradient computation is required.


Memorized Training Sample
"Living in the Light with Ann Graham Lotz"
Generation with No Blocked Neurons
Initial Selection
Identify Candidate Neurons With Out-of-Distribution Activations
Candidate Neurons
Neuron#124
Neuron#221
Neuron#362
Generation with Blocked Candidate Neurons
Refinement
No Blocked Neurons
Noise Similarity
Blocked Candidates
Remove Candidate Neurons Not Responsible for Memorization
Memorizing Neurons
~~Neuron#124~~
Neuron#221
~~Neuron#362~~
Generation with Blocked Memorizing Neurons

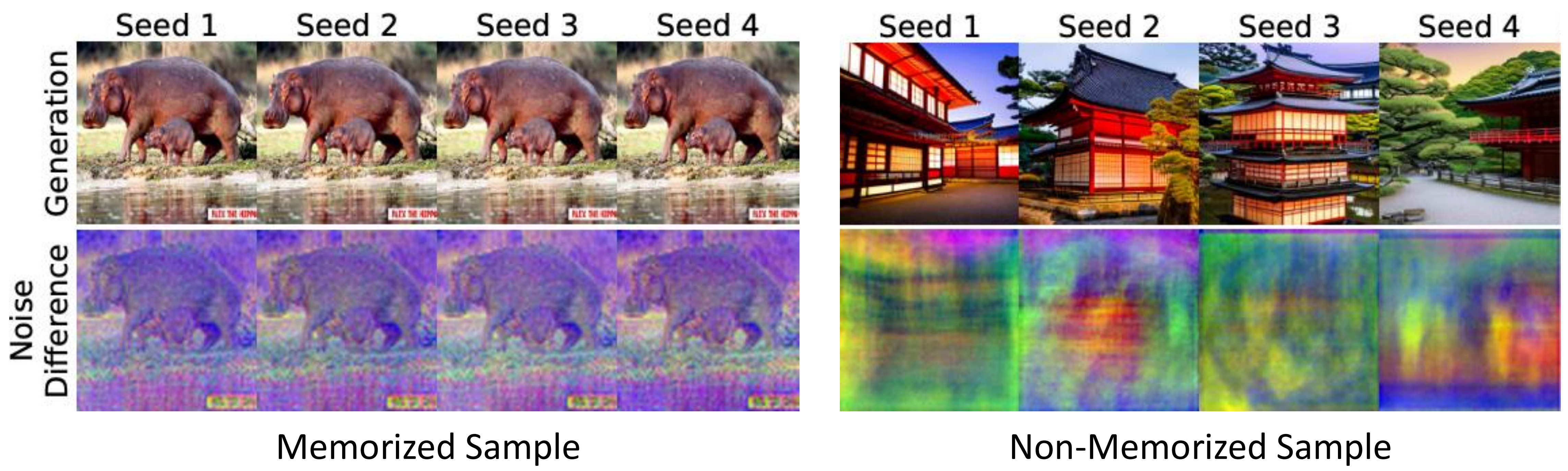
Quantifying Memorization Strength

Memorization strength is quantified by measuring the **similarity between the denoising trajectories** starting from different initial noises.



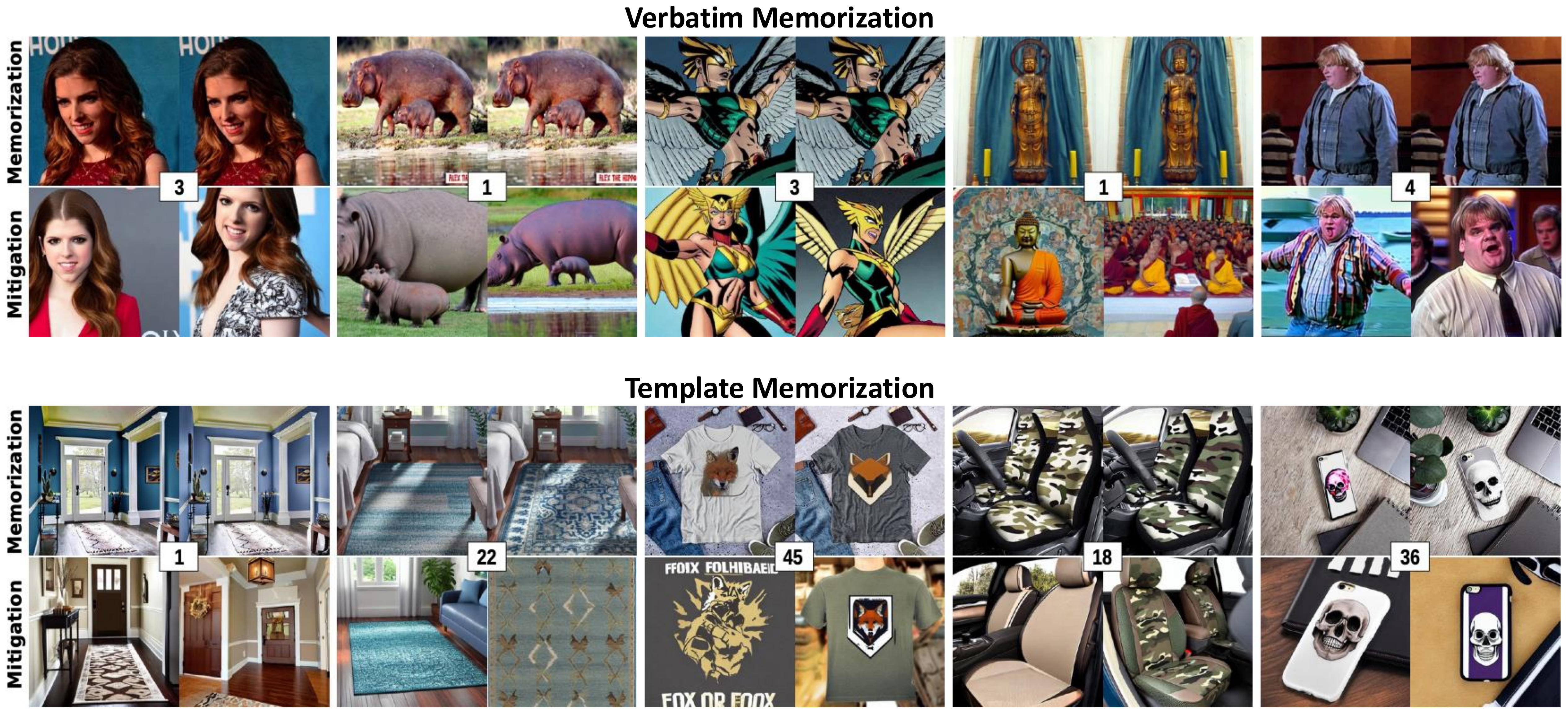
Noise Differences

The denoising **trajectories are consistent** for memorized samples but vary substantially for non-memorized content.



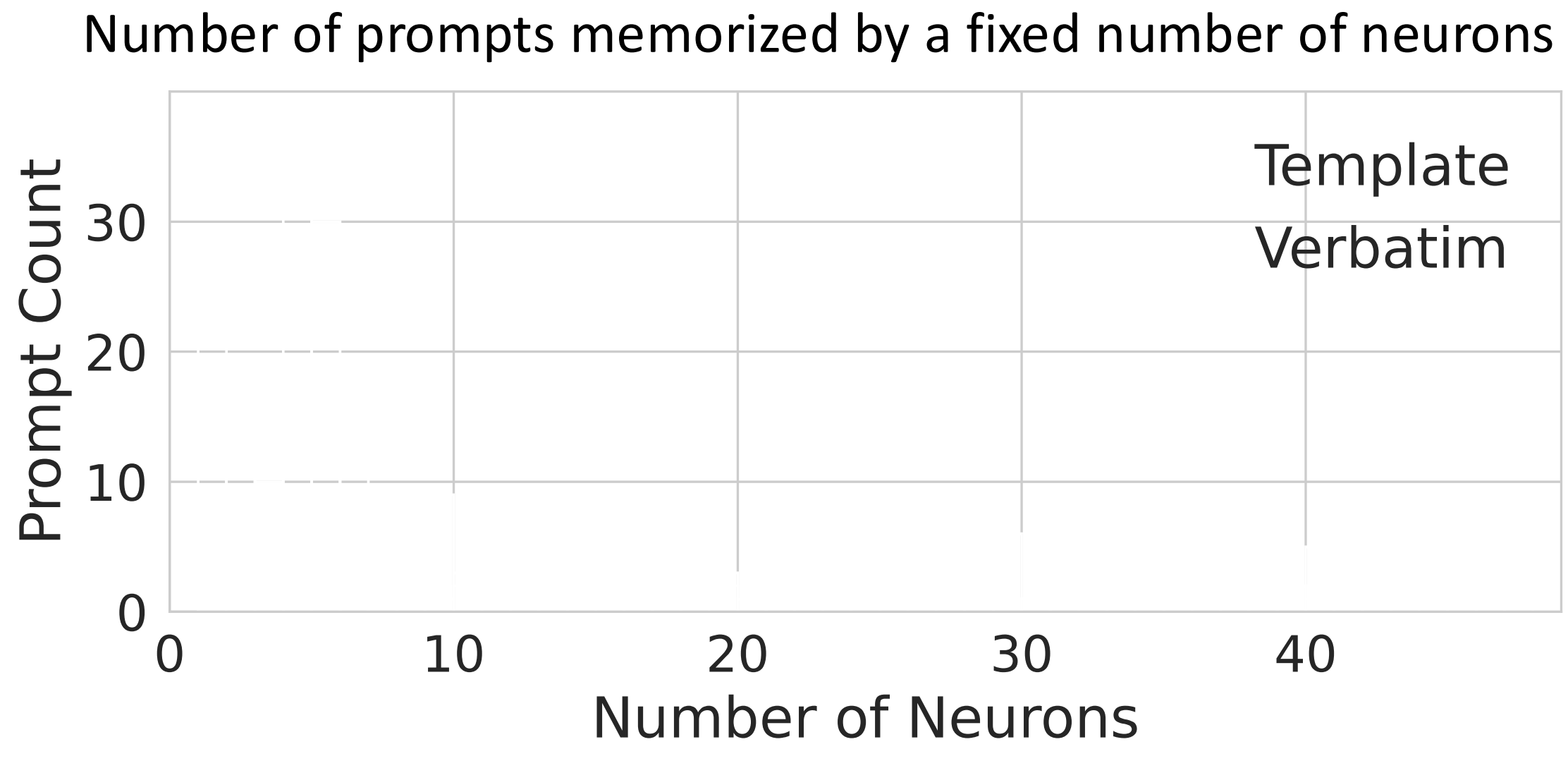
Effect of Deactivating Memorization Neurons

Deactivating memorization neurons **increases diversity** and **mitigates memorization**. Only a **few neurons** are responsible for memorization.



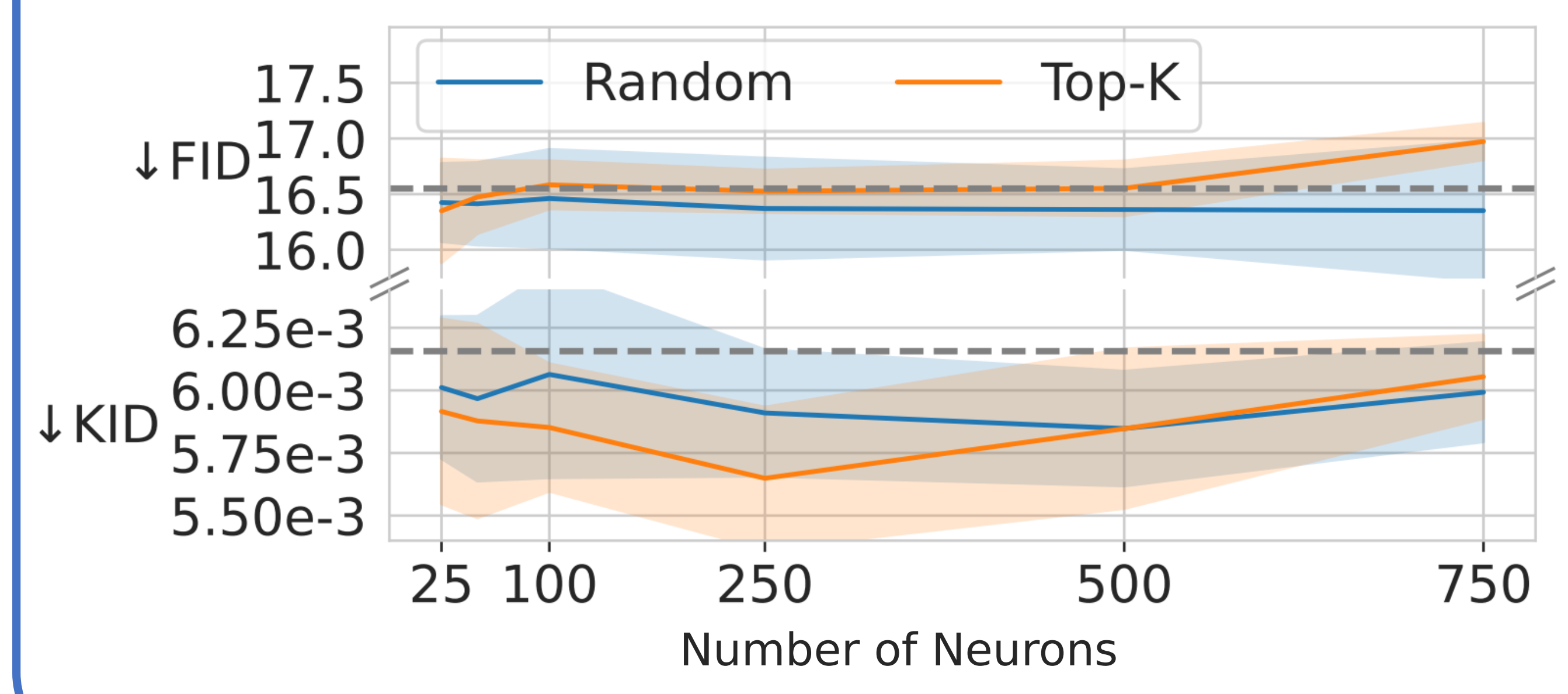
Distribution of Memorization Neurons

A **small set of neurons** is responsible for memorization.



Quality Retention

Deactivating memorization neurons **does not degrade image quality**.



Code & Paper

Contact

Please feel free to reach out to us!

Dominik Hintersdorf

✉ dominik.hintersdorf@dfki.de

𝕏 @d_hintersdorf

🌐 d0mih.github.io

Lukas Struppek

✉ lukas.struppek@dfki.de

𝕏 @LukasStruppek

🌐 lukasstruppek.github.io