IT UNIVERSITY OF COPENHAGEN

CVPR Nashville June 11-15, 2026

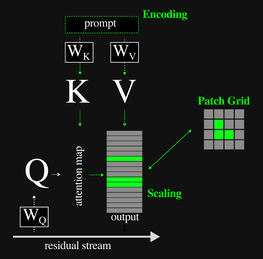
Northeastern University

# Patch Explorer

## An Interface for Interpreting Diffusion Models through Interaction

Imke Grabe*, Jaden Fiotto-Kaufman*, Rohit Gandikota, David Bau

**Patch Explorer is an interface for interacting with the hidden states of diffusion models by directly manipulating the cross-attention heads through interventions.**



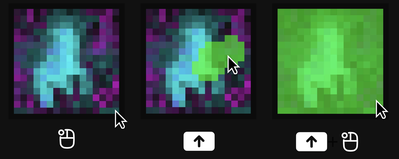

A **Patch Grid** represents the output of an attention heads spatially.

**Scaling** multiplies the output of the attention head by a given factor.

**Encoding** replaces it with the output for an alternative text encoding.

A user can interact with patch grids by:

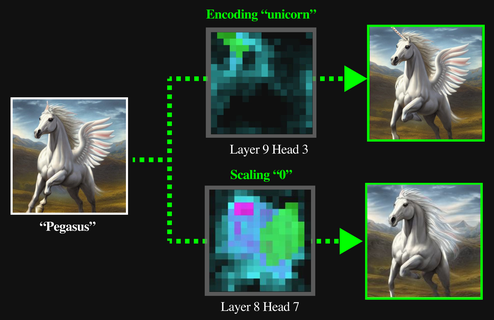
1. choosing the intervention
2. targeting timesteps
3. selecting patches



Choosing a patch grid. Drawing to select patches. Selecting all patches at once.

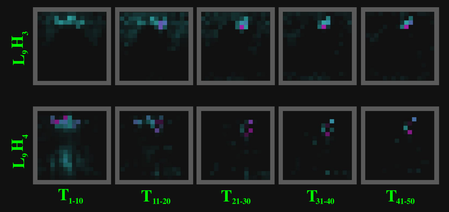
### Drawing with patches

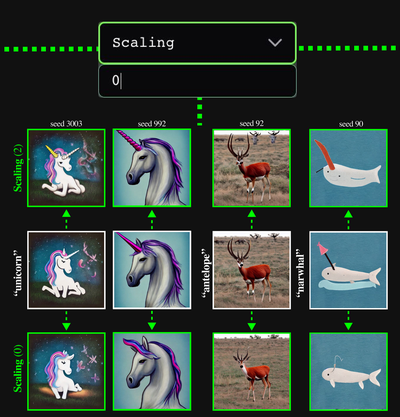
By clicking on a patch grid, a user can select individual patches by moving the mouse while holding Shift down, which marks them green. Clicking selects all patches at once.




### Finding Semantic Features in Attention Heads

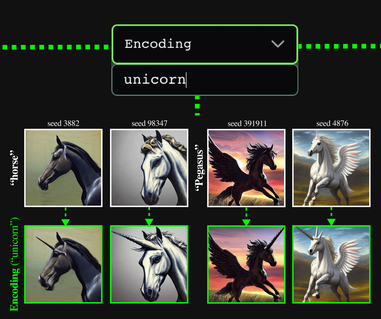
By inspecting patch grids over timesteps, the user can find patches that light up specifically for the unicorn's horn, for example Head 3+4 at Layer 9.



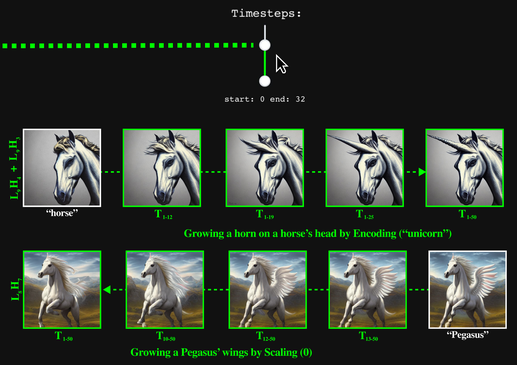

### Validating Feature Attribution

Scaling the attention heads enhances or removes the feature. The same intervention also alters other concepts, like an antelope's antlers or a narwhal's horn.




### Transferring Features across Concepts

By encoding 'unicorn' at the attention heads, the feature can be transferred to horses and Pegasi.




### Tracing Features across Timesteps

Intervening on certain timestep ranges shows how features form.

**More examples, paper, code and demo:**

patch.baulab.info

built with NN