High-resolution image reconstruction with latent diffusion models from human brain activity

Yu Takagi and Shinji Nishimoto

CVPR 2023

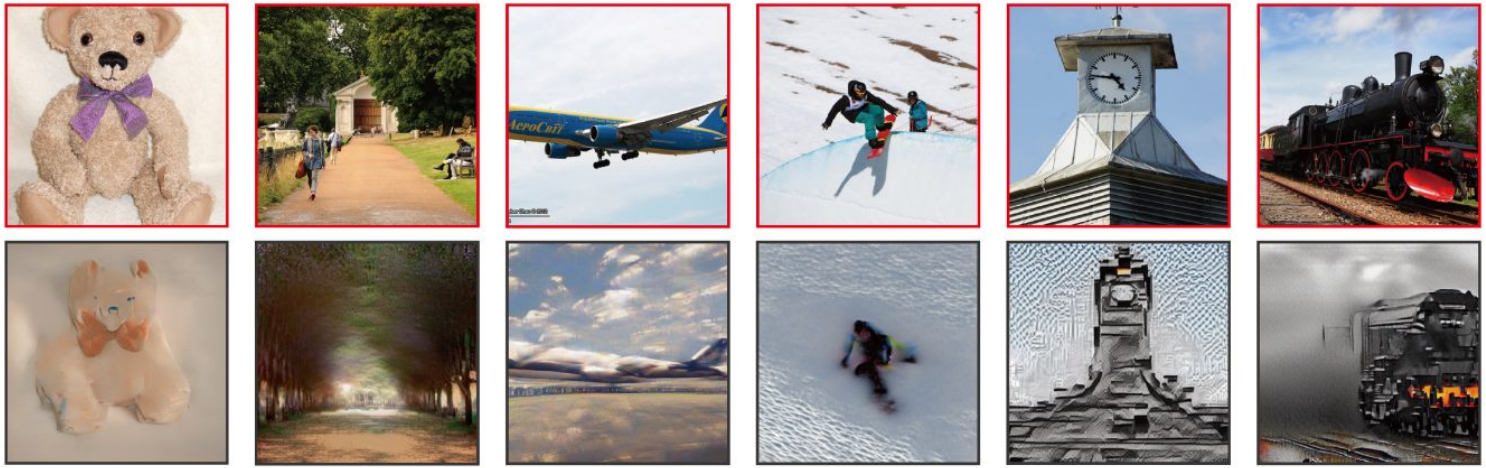
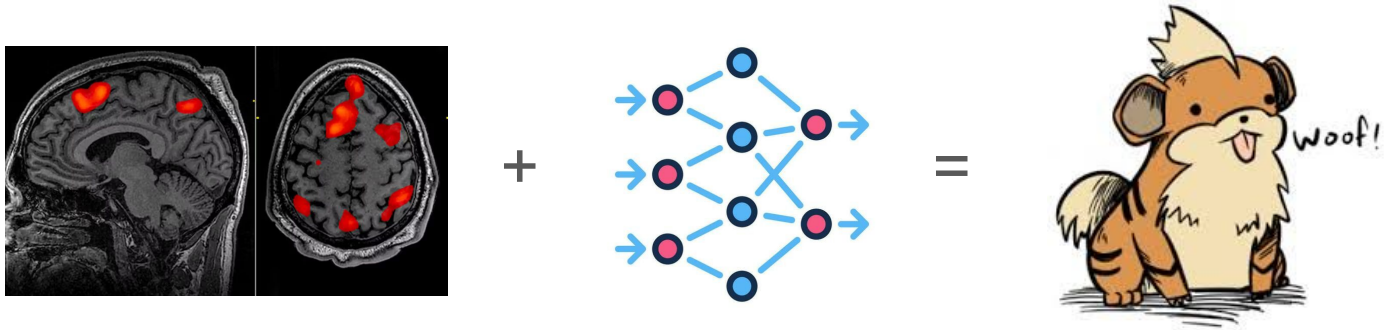


Figure 1. Presented images (red box, top row) and images reconstructed from fMRI signals (gray box, bottom row) for one subject (subj01).

Introduction

Understanding how the human brain processes and represents visual information is a fundamental challenge in both neuroscience and artificial intelligence.

This study introduces Latent Diffusion Models (LDMs)—a more efficient variant of diffusion models—to reconstruct high-resolution images directly from fMRI signals, without additional model training.



Motivation

- **Enhancing Visual Reconstruction** – Previous methods required large datasets and fine-tuning, limiting scalability.
  - LDMs generate high-resolution images with strong semantic accuracy while reducing computational costs.

- **Bridging AI and Neuroscience** – Understanding how LDMs process information helps reveal similarities between artificial vision models and human brain activity.
  - This work not only improves image reconstruction but also provides a biological interpretation of diffusion models.

Contributions

1. **High-Resolution Image Reconstruction**
   a. Demonstrate that a simple framework can reconstruct 512×512 images from fMRI signals with high semantic fidelity.
   b. Unlike previous methods, this approach requires no additional training or fine-tuning of deep generative models.
2. **Neuroscientific Interpretation of LDM Components**
   a. Provide a quantitative analysis of how different components of an LDM correspond to distinct brain regions.
   b. This mapping helps bridge the gap between AI-based generative models and biological vision systems.
3. **Understanding Text-to-Image Processing in LDMs**
   a. Present an objective interpretation of how the LDM's text-to-image conversion process balances semantic information from textual conditioning while preserving the visual appearance of the original image.

Dataset: Natural Scenes Dataset (NSD)

- **Key Features of NSD**
  - Collected using a 7-Tesla fMRI scanner over 30–40 sessions per subject.
  - Subjects viewed three repetitions of 10,000 natural images from MS COCO (cropped to 425×425 pixels).
  - Data includes voxel-wise fMRI activity across multiple brain regions.
- **Dataset usage in this study**
  - 4 subjects were analyzed (subj01, subj02, subj05, subj07).
  - 27,750 trials used for training (the three separate trials without averaging were used); 2,770 trials for testing (the average of the three trials associated with each image was used).
  - fMRI signals from early and higher visual cortices were mapped to Latent Diffusion Model (LDM) components.

Previous Works on Reconstructing visual images from fMRI

- **Traditional Feature-Based Methods**
  - Early studies used handcrafted features to reconstruct images from fMRI data.
  - These methods relied on low-dimensional representations, such as edge detection or spatial frequency analysis, but lacked the ability to capture complex visual details.
- **Deep Learning-Based Approaches**
  - The emergence of deep neural networks (DNNs) improved reconstruction by learning hierarchical visual features from brain signals.
  - Approaches using Generative Adversarial Networks (GANs) and self-supervised learning enabled better visual fidelity, but still suffered from : Low resolution (typically ≤256×256 pixels), the need for large-scale training on fMRI datasets and limited generalizability across subjects and stimuli.

Previous Works on Reconstructing visual images from fMRI

- **Use of Semantic Information for Improved Fidelity**
  - Some studies introduced textual or categorical labels as additional input to enhance semantic accuracy.
  - However, these methods still required fine-tuning of generative models and lacked the flexibility to generalize across different types of images.

Limitations with earlier approaches

- Small sample sizes in neuroscience make training complex models challenging.
- Many approaches require extensive fine-tuning, limiting their adaptability.
- Achieving high-resolution and semantically accurate reconstructions remains a significant challenge.

Why use Latent Diffusion Models?

- State-of-the-Art Performance in Image Generation
  - Diffusion Models (DMs) have achieved cutting-edge results in various tasks, including conditional image generation, super-resolution, and image colorization.
- Efficiency and Scalability with LDMs
  - LDMs improve upon traditional DMs by operating in a compressed latent space, reducing computational costs while maintaining generative performance.
- High-Resolution and Semantic Fidelity
  - LDMs excel at generating high-resolution images (512×512 pixels) with strong semantic coherence, making them well-suited for brain-to-image translation.

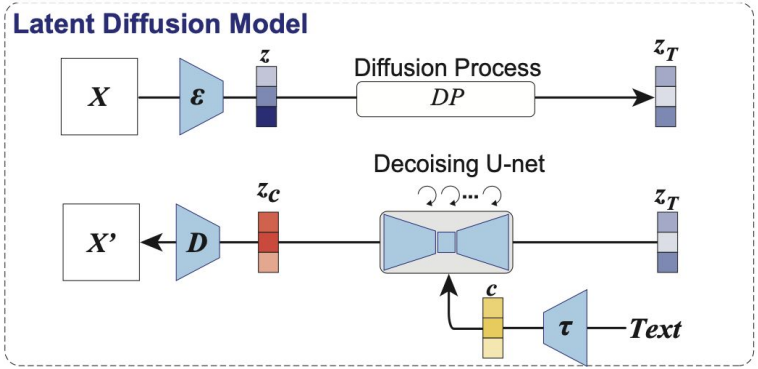
Why use Latent Diffusion Models?

- A New Framework for Understanding Generative Models
  - Despite their success, LDMs remain poorly understood, particularly in how they process latent signals and how noise impacts image generation.
  - By applying LDMs to brain decoding, this study not only advances visual reconstruction but also sheds light on the internal workings of these powerful generative models.

Latent Diffusion Models

- Latent Diffusion Models (LDMs) are a class of deep generative models that generate high-quality images by progressively refining noisy inputs.

- They improve upon traditional Diffusion Models (DMs) by operating in a lower-dimensional latent space, making them computationally more efficient.



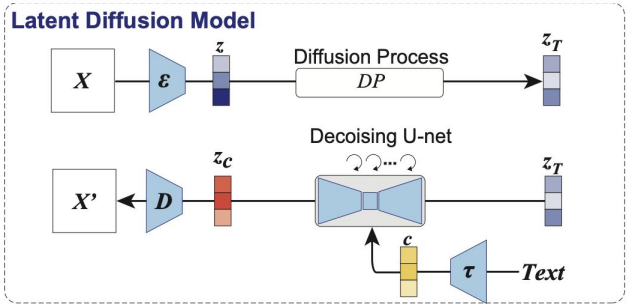
Working of Latent Diffusion Models

1. **Latent Space Encoding**
   a. Instead of working in pixel space, LDMs compress images into a lower-dimensional latent space using an autoencoder.
   b. This step reduces computational cost while preserving essential image details.

2. **Diffusion Process (Adding Noise)**
   a. The model gradually adds noise to the latent representation of an image, making it increasingly unrecognizable.
   b. This trains the model to learn how images degrade over time

3. **Reverse Diffusion (Denoising Process)**
   a. During generation, the model reverses the noise-adding process step by step, reconstructing the image from pure noise.
   b. A U-Net architecture helps predict and remove noise at each step.

4. **Conditioning for Controlled Generation**
   a. LDMs allow control over image generation using conditioning inputs (e.g., text descriptions, fMRI signals).
   b. This enables text-to-image synthesis and brain-to-image reconstruction.



$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon_t$$

Decoding: reconstructing images from fMRI

This study uses Latent Diffusion Models (LDMs) to reconstruct high-resolution images from fMRI signals, employing a simple yet effective three-step process which efficiently maps brain activity to visual representations, reconstructing high-resolution, semantically meaningful images from fMRI.

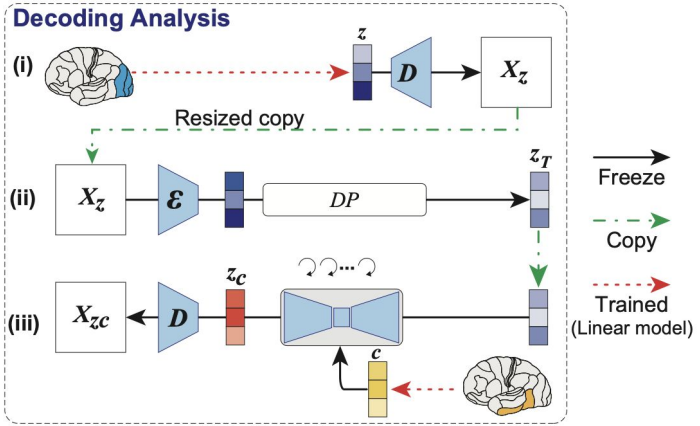
Decoding: reconstructing images from fMRI

The visual reconstruction from fMRI signals was performed using LDM in three steps:

1. **Predicting Latent Representations from fMRI**
   a. fMRI signals from the early visual cortex are used to predict the latent representation (z) of the viewed image.
   b. The predicted z is then passed through the LDM's decoder, generating a coarse reconstruction (Xz) of the image.

2. **Noise Addition and Latent Space Processing**
   a. The reconstructed image (Xz) is encoded back into latent space and corrupted with noise through the diffusion process.
   b. This step ensures alignment with the structure of the LDM, improving the final reconstruction.

3. **Refining Reconstruction Using Text Representations**
   a. fMRI signals from the higher visual cortex are used to decode text-related latent representations (c).
   b. The LDM combines the noisy latent representation (zT) with the decoded text representation (cc) to refine the image, generating the final reconstruction (Xzc).



Encoding: Whole-brain Voxel-wise Modeling

The paper explores how different components of Latent Diffusion Models (LDMs) correspond to brain activity. To achieve this, whole-brain voxel-wise encoding models are built to predict fMRI responses from LDM features.

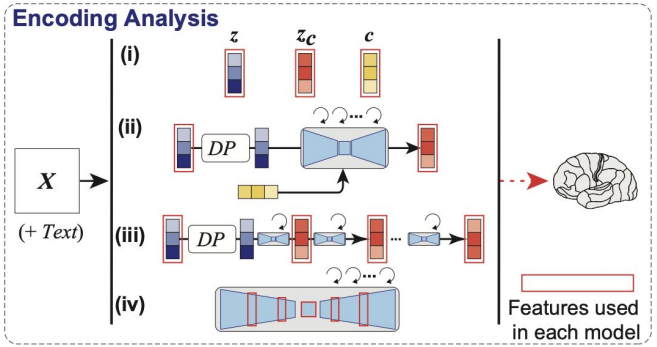
Encoding: Whole-brain Voxel-wise Modeling

1. **Mapping Brain Activity to LDM Components**
    a. Linear encoding models are trained to predict voxel-wise fMRI signals using the following LDM components:
        i. z – Latent representation of the original image (linked to early visual cortex).
        ii. c – Latent text representation (linked to higher visual cortex).
        iii. zc – Final image representation after text conditioning.

2. **Investigating the Impact of Noise Levels**
    a. The study examines how varying noise levels in zc affect brain activity predictions.
    b. Lower noise retains original image details, activating early visual regions, while
    c. Higher noise strengthens semantic content, activating higher visual areas.

3. **Analyzing the Denoising Process in LDMs**
    a. By extracting intermediate LDM representations at different denoising steps, the study reveals:
        i. Early denoising stages correlate with low-level visual processing (V1/V2).
        ii. Later denoising stages align with higher-order regions (ventral visual stream).

4. **Layer-Wise Interpretation of U-Net in LDMs**
    a. Encoding models are built for different layers of U-Net within Stable Diffusion.



Evaluation

- Model weights were estimated from training data using L2-regularized linear regression, and subsequently applied to test data.

- For evaluation, Pearson's correlation coefficients was used between predicted and measured fMRI signals. The statistical significance (one-sided) was computed by comparing the estimated correlations to the null distribution of correlations between two independent Gaussian random vectors of the same length (N=982). The statistical threshold was set at $P < 0.05$ and corrected for multiple comparisons using the FDR procedure.

- All feature dimensions were reduced to 6,400 by applying principal component analysis, by estimating components within training data.

Results: Image Reconstruction from fMRI

- **Using only latent representation (z):**
  - Images are **visually consistent** with the original.
  - Lacks **semantic accuracy** (e.g., shape and object identity).

- **Using only text-based semantic representation (c):**
  - Captures **high-level meaning** of the original image.
  - Fails to **preserve fine details** and visual consistency.

- **Using both latent and semantic representations (zc):**
  - Achieves **high-resolution and high semantic fidelity**.
  - Best balance between **appearance and meaning**.

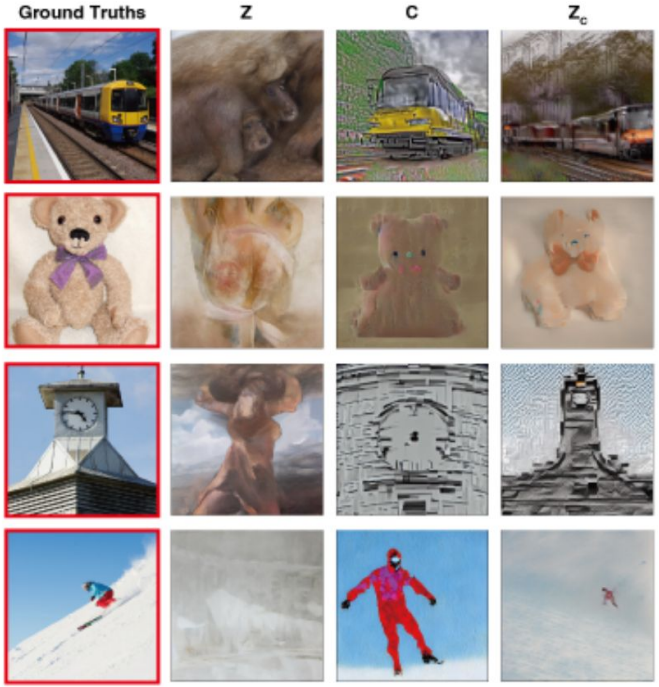
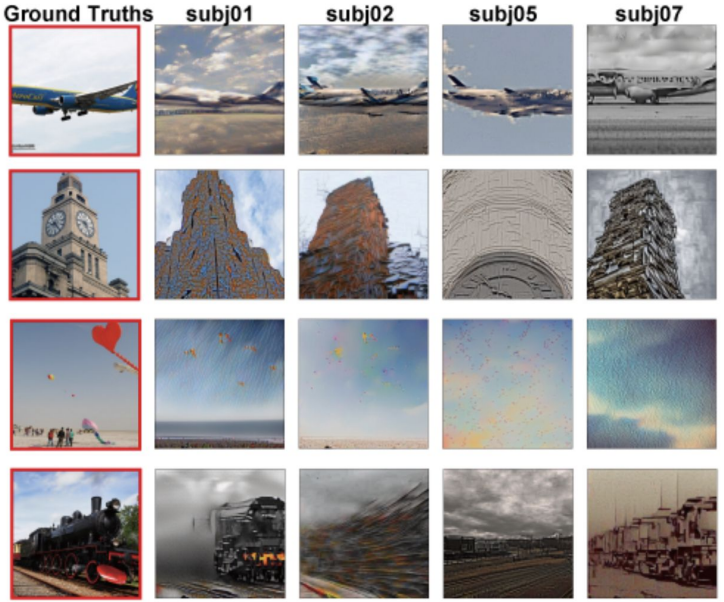


Figure 3. Presented (red box) and reconstructed images for a single subject (subj01) using $\mathbf{z}$, $\mathbf{c}$, and $\mathbf{z_c}$.

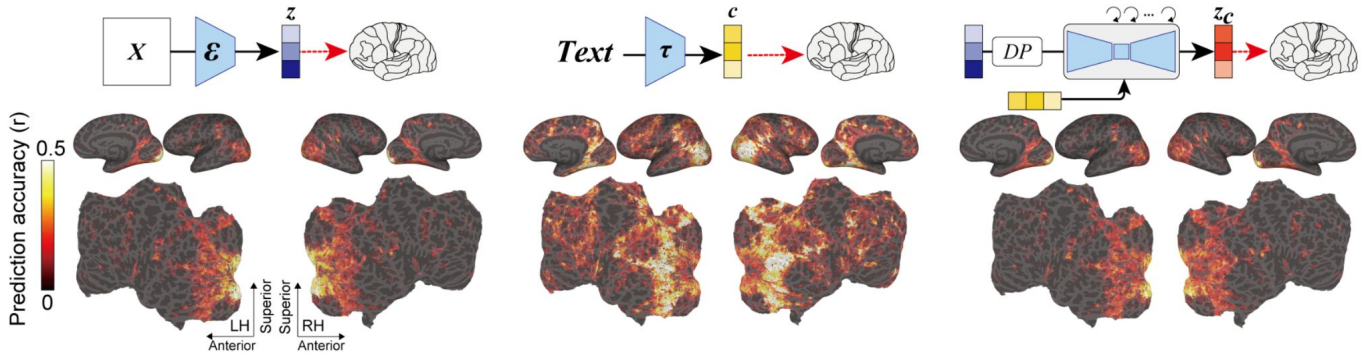
Subject Variability in Reconstruction

- Reconstruction quality varies across subjects due to differences in brain activity patterns and data quality.
- Subjects with higher fMRI decoding accuracy produced more detailed and semantically accurate reconstructions.
- Variability may reflect individual differences in perception, neural representation, or fMRI signal strength.
- Despite differences, consistent brain-LDM correspondences were observed across subjects.



Encoding: Comparison among Latent Representations

- **z (Latent Representation):**
  - High prediction in **Early Visual Cortex** (posterior region).
  - Some activation in **Higher Visual Cortex** (anterior region).
- **c (Semantic Representation):**
  - Highest prediction in **Higher Visual Cortex** (semantic processing).
- **zc (Combined Representation):**
  - Similar to **z**, strong in **Early Visual Cortex**.
  - **Reducing noise** in zc makes it closer to z.

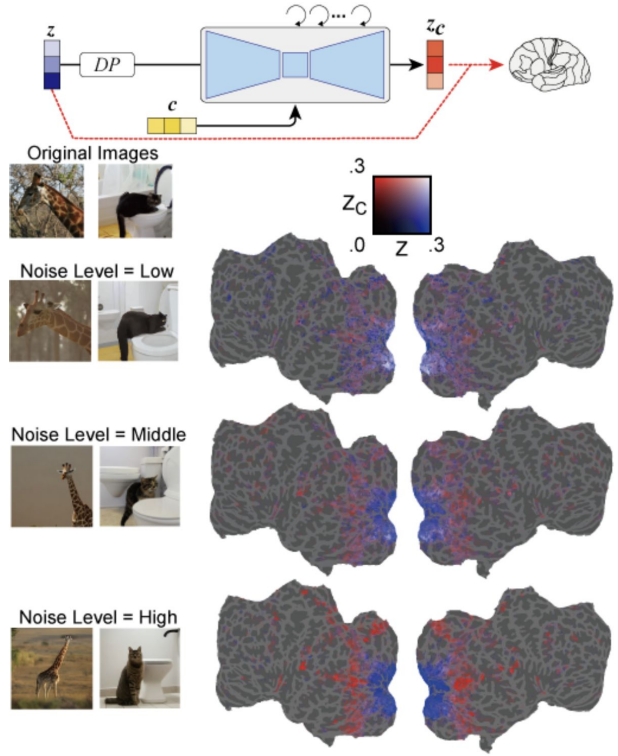


Encoding: Comparison across different noise levels

z vs. zc Prediction Accuracy:

• With **low noise**, **z** predicts voxel activity better across cortex.

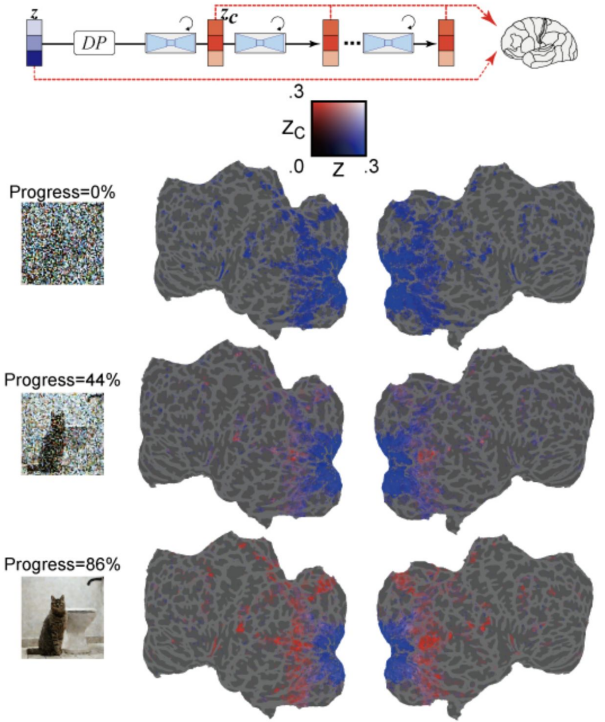
• With **higher noise**, **zc** shows stronger activation in **Higher Visual Cortex** (semantic processing).



Encoding:Comparison across different diffusion stages

• **Early Stage: z** dominates fMRI prediction (**low-level visual details**).

• **Middle Stage: zc** shows stronger activation in **Higher Visual Cortex** (**semantic content emerges**).

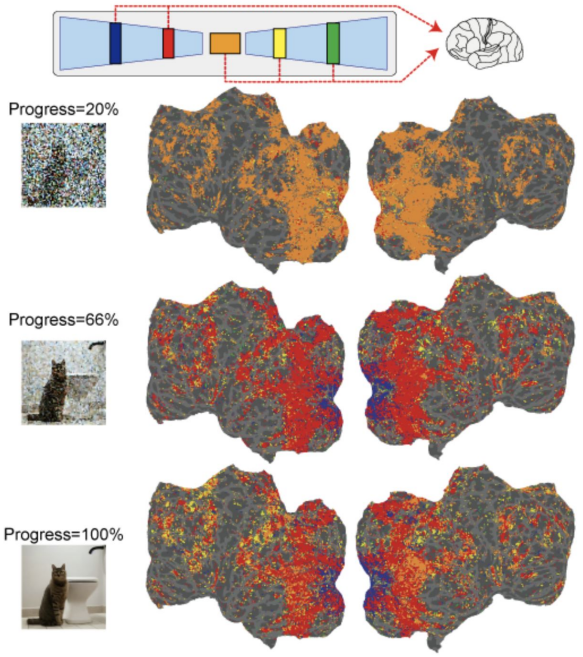


Encoding: Comparison across different U-Net Layers

Early Phase: Bottleneck layer has the highest prediction across cortex (**compressed image information**).

Later Phase:

- **Early U-Net layers** predict activity in **Early Visual Cortex** (fine details).
- **Bottleneck layer** shifts to **Higher Visual Cortex** (semantic processing).



Why differentiate early vs. higher visual areas of the cortex?

Understanding the distinction between early and higher visual areas helps reveal how the brain processes and reconstructs images, aligning with AI models like Latent Diffusion Models (LDMs).

1. **Early Visual Areas (V1/V2) – Low-Level Processing**
    a. Located in the occipital lobe, first to receive visual input.
    b. Encodes basic features like edges, contrast, orientation, and motion.
    c. Strongly retinotopic (spatially maps visual input).
    d. Aligns with LDM's latent representation z, which captures raw image structure.

2. **Higher Visual Areas (V4, IT, Ventral Stream) – High-Level Processing**
    a. Process complex features, including object recognition, textures, and semantic meaning.
    b. Less retinotopic, focusing on global image interpretation.
    c. Linked to LDM's text-conditioning (c), which refines images with semantic information.

This functional distinction aligns with the two components of the LDM (low-level latent z vs. semantic textual embedding c).

Findings

- **High-Resolution Brain-to-Image Reconstruction**
  - Successfully reconstructed 512×512 images from fMRI signals with high semantic fidelity.
  - Achieved state-of-the-art performance without additional deep-learning model training or fine-tuning.
- **Brain-LDM Correspondence**
  - Early visual cortex encodes low-level visual features (linked to LDM's latent image representation z).
  - Higher visual cortex encodes semantic information (aligned with LDM's text representation c).
- **Understanding the Denoising Process in LDMs**
  - Early denoising stages correspond to low-level brain areas (e.g., V1/V2) (Primary Visual Cortex /Secondary Visual Cortex)
  - Later denoising stages align with higher-order visual regions (ventral visual stream).
- **U-Net Layer Correspondence to Brain Activity**
  - Early U-Net layers match fine-grained visual processing in early brain regions.
  - Deeper U-Net layers capture semantic information, linked to higher-order cortical areas.

Conclusion

- **Proposed a novel visual reconstruction method using LDMs which can**
  - Reconstruct high-resolution images with high semantic fidelity from human brain activity
  - Does not require training or fine-tuning of complex deep-learning models
  - Only requires simple linear mappings from fMRI to latent representations within LDMs

- Provided a quantitative interpretation for the internal components of the LDM by building encoding models

- **Limitations and Future Scope:** The internal processes of DMs remain poorly understood, and this study is the first to provide a quantitative interpretation from a biological perspective.

Thank You.