Brain Imaging and Computational Models: Visual Information Reconstruction in Neural Networks

Computational Models in Cognitive Science - 10944

**Instructor:** Noam Shental

**Author:** Itamar Yeger

**ID Number:** 31355183

**Phone Number:** +972-050-4530021

The Open University of Israel

**Date:** 26/11/2024

Brain Imaging and Computational Models: Visual Information Reconstruction in Neural Networks

|  |  |  |
| --- | --- | --- |
| Table of contents | | |
| **1.** | **Introduction** | **3-6** |
| 1.1 | Visual Recognition and Neural Activity | 3-5 |
| 1.2 | Brain imaging and fMRI | 5-6 |
| **2.** | **Stable diffusion** | **6-16** |
| 2.1 | Mathematics | 7-10 |
| 2.2 | Algorithm | 10-14 |
| 2.2.1 | Forward process | 10-11 |
| 2.2.2 | Reverse process | 11-14 |
| 2.3 | Application for generation | 14-16 |
| **3.** | **Visual Information Reconstruction in Neural Networks** | **16-26** |
| 3.1 | Results | 20-24 |
| 3.2 | Biological interpretation | 25-26 |
| **4.** | **Discussion** | **27-28** |
| **5.** | **References** | **29-30** |

1. **Introduction**

This work presents a new approach to reconstructing high-resolution images from brain activity. The goal of this seminary paper is to explain and simplify the interactions between brain imaging, visual perception, and computational models for those not familiar with the subject. To achieve this, the basic concepts of the process of perception and visual processing in the brain will be explored therefore, topics such as how visual information is processed, the brain areas involved, and how an image is formed in our minds will be discussed. We will put particular attention on the discussion on brain imaging techniques with an emphasis on functional magnetic resonance imaging (fMRI) and how it helps in reconstructing subjective images.

Next, we will discuss the Stable Diffusion Model. a deep learning model that has gained attention in recent years. This model is prominently used in OpenAI's DALL\_E project to create high-resolution images from textual descriptions (Figures 1 and 2). The model operates by gradually adding random noise to a specific object in a process of resembling particle diffusion, hence its name. A neural network then learns to reconstruct the original object from this noise [1], like an artist carving a figure from wood.



**Figure 2: An illustration created by DALL-E based on the text "Broccoli jumping on a trampoline and reading a book."**



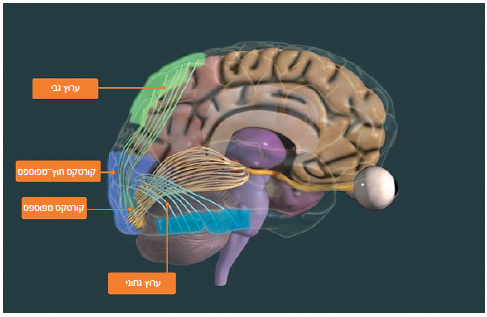
**Figure 1:** An illustration created by DALL-E based on the text "Astronaut surfing on the Milky Way."

We will also examine the study "High-resolution image reconstruction with latent diffusion models from human brain activity" by Yu Takagi and Shinji Nishimoto (2023). This study combines visual perception, brain images and computational models using Latent Diffusion Models to reconstruct high-resolution images from fMRI data of human subjects [2]. Finally, we will discuss the cognitive, computational, and philosophical implications of the study: the success of such research and others in the field could contribute to a deeper understanding of the relationship between brain activity and visual perception.

* 1. **Visual Recognition and Neural Activity**

To understand our environment and behave appropriately we must quickly and accurately determine what objects are around us. This means that the visual system must create a mental representation of a given object. The term "mental representation" refers to a pattern of neural activity in the brain that contains information about stimulus and creates a subjective perceptual experience of the stimulus. In the visual system, the process of creating representations begins the moment light receptors in the retina start responding to light, forming representations containing information about the spatial pattern of intensity and wavelength in the retinal image, as neural signals flow to the primary visual cortex (V1) and then to higher visual area, more complex representation of the retinal image is constructed [3].

As shown in figure 3, several brain areas are involved in this visual process. The retina converts light into electrical signals, transmitted through the optic chiasm to the lateral geniculate nucleus (LGN) in the thalamus, which relays visual information to the rest of the brain: to the primary visual cortex (V1) and then to the visual association cortex (V2) [4].



**Figure 3 (adapted from [4], p. 230):** Brain regions essential for vision: striate cortex, extrastriate cortex, dorsal stream, and ventral stream.

Lets review those brain areas:

**Lateral Geniculate Nucleus (LGN)** – After the optic chiasm, visual signals reach the LGN, organized into six layers that all receive feedback from higher visual areas. The LGN distinguishes between two systems that receive input from different types of cells in the retina. These systems are responsible for analyzing different types of visual information: one system is sensitive to movement, depth, and brightness differences, while the other specializes in color and shape analysis.

**Primary Visual Cortex (V1)** – Located in the posterior occipital lobe, receiving the main input from the visual system. Most researchers believe the cortex is organized into modules – components in the brain responsible for performing specific tasks with unique input and output. Each module contains hundreds of thousands of neurons, receiving information from other modules, performing computations, and transmitting the results.

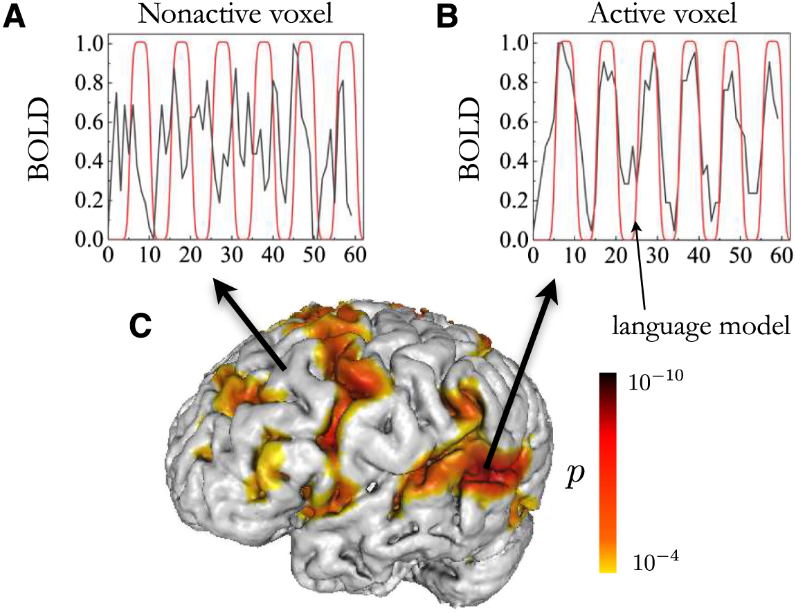
**Visual Association Cortex (V2)** – This area receives information, among other, from the primary visual cortex and projects to the inferior temporal cortex. Integration of information from all modules and pathways occurs in the extrastriate cortex. The cortex includes several areas, each specializing in a specific aspect of visual information, such as direction movement, color, etc. Visual processing in the extrastriate cortex splits into two pathways: the dorsal stream, also called the "where" pathway, specializing in location and movement recognition, and the ventral stream, also called the "what" pathway, reaching the temporal lobe specializing in identifying the meaning of objects. Therefore, to determine what objects are around us, the visual system must create a mental representation of a given object, and it does so through the activity of various organs and structures in the brain that transmit information to each other, each with different roles and responsibilities. The work of all these components tighter creates the mental representation – the image we perceive as reality [3, 4, 17].

* 1. **Brain Imaging and fMRI**

Physiological, anatomical, and cognitive studies indicate that the brain has different processing areas anatomically. It can be seen in the fact that during cognitive tasks, local changes in neural activity cause local changes in metabolism, blood flow, and blood volume. So, these changes can be used to map the functional loci of components in some mental activities [5].

fMRI (Functional Magnetic Resonance Imaging) is a technique that measures changes in the oxygen level in blood, indicating brain areas active during different cognitive tasks or brain states. If so, fMRI is used to investigate BOLD (blood-oxygen-level-dependent) activation in the human brain for both clinical and research purposes. Although fMRI cannot fully resolve the issue of "functional specialization" of brain areas by itself, it helps shed light on which areas are at least involved in certain cognitive processes [6].

Functional imaging techniques provide tools to explore the function of the human brain. An example has been demonstrated by Li et al. (2020), Where subjects' brains were scanned using fMRI while they performed verbal tasks using verb creating in response to auditory nouns. During the verb creation task, subjects were presented with a noun (e.g., baby) verbally and then asked to create action words (e.g., crying, crawling) related to the noun. That is using fMRI to map the functional architecture of the brain, focusing on language processing.



**Figure 4 (adapted from [6], Fig. 1):** Brain activity map of a subject. BOLD signal for inactive and active voxels shown in A and B, respectively. A threshold level for individual voxels (p < 0.0001) was set to minimize random noise fluctuations.

A voxel is a term describing the smallest unit of a 3D data, analogous to a pixel in a 2D image. Instead of representing a point on a surface, a voxel represents a small volume within the data (the brain). Each voxel contains information about the properties of that part of the brain, such as density or neural activity, allowing researchers to examine the brain with high precision. Different colored voxels can be seen on the brain in figure 4 where we can see the brain activity of an individual performing a language task. Gray areas represent inactive voxels (A), and areas where active voxels are represented by a color scale from yellow to black to indicate the level of activity (B). Active voxel areas are specific brain regions like Broca's area, Wernicke's area, the inferior frontal gyrus (IFG), and the temporal lobe, all involved in language comprehension or production [6].

In addition to identifying brain activity, efforts are underway to use fMRI data to reconstruct the perceptual stimulus causing specific brain activity. This means understanding what influenced the brain to generate such a pattern of activity. In a study by Kay et al. (2008), researchers recorded fMRI data while subjects viewed natural images to estimate models for each voxel. They then used these models to predict brain activity when subjects viewed new images, identifying the specific image based on the predicted activity [7].

Building on this foundation, Takagi and Nishimoto (2023) employ an innovative methodology, utilizing Stable Diffusion—a generative model—to reconstruct high-resolution images from fMRI data. This method could enhance our ability to visualize and interpret complex brain functions.

1. **Stable Diffusion:**

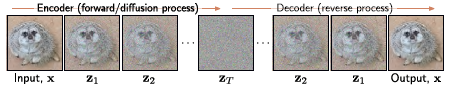
The Stable Diffusion model employs a stochastic process known as a Markov chain, where each state depends only on the previous state. This approach is crucial in various fields, including simulating complex systems and machine learning. The method iteratively transitions from one distribution to another, an idea used in non-equilibrium thermodynamics, effectively tracing, and reconstructing lost information.

In Stable Diffusion, this process begins by adding Gaussian noise to the original data. Over several steps, the model progressively corrupts the data with more noise, simulating a forward diffusion process. The reverse diffusion process then comes into play: the model learns to iteratively remove the noise, gradually transforming the noisy data back into the original data distribution. This is akin to tracing back a droplet of liquid diffused into a large container of water.

By progressively refining the reconstruction through the iterative removal of noise, the diffusion process allows the model to generate new data that is statistically similar to the original data [8].

Stable Diffusion can be seen as having two processes: an encoder (Forward process) and a decoder (Reverse process).

The encoder takes sample data X (e.g., an image) and maps it through a series of intermediate variables Z1...ZT by adding Gaussian noise. Figure 5 shows the process illustrated with an owl image. The assumption is that with enough steps, the conditional distribution q (ZT |x) and the marginal distribution q(ZT ) of the final variable both become standard normal distributions. This means that regardless of the initial data point X, after the diffusion process, the final representation (ZT) will follow a normal distribution independent of X. This characteristic is crucial because it simplifies the reverse mapping in the decoding process. Since this process is predefined, all learned parameters are in the decoder [9].



**Figure 5 (adapted from [9], Fig. 18.1):** Diffusion model. The encoder maps x through a series of variables and adds noise until zT. The decoder learns how to transfer the information back through the latent variables while "removing" noise at each stage.

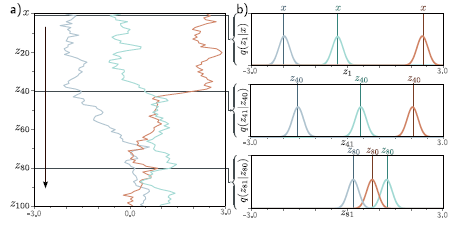
The decoder reverses the encoding process: it starts with ZT and maps back through ZT−1...z1 until the data point X is reconstructed. The reverse encoding process occurs with a series of neural networks trained to map backward between each pair of neighboring latent variables Zt and Z−1. The result is that noise is gradually removed from the representation until a realistic data sample remains [9].

* 1. **Mathematics**

**Encoding:** As mentioned, the diffusion process starts with a data sample X. It then passes the sample through a series of intermediate variables Z1, Z2...ZT, each the same size as X. This process follows the equations:

* ϵt=random noise sampled from a standard normal distribution with mean μ and standard deviation σ.
* βt=a hyperparameter in the range [0,1] determining the amount of noise added at ecah step t. β is a vector with T elements, each between 0 and 1.

The probability Zt depends only on the value immediately before, the variable Zt-1. With enough stages (T), it can be said that traces of the original data disappear, and q(Zt|X) = q(Zt) becomes a standard normal distribution [9].



**Figure 6 (adapted from [9], Fig. 18.2)**: Encoding process (forward process)

**(a)** One-dimensional data x with T=100 variables z1...z100 and β=0.03 at all stages. Three values of x (gray, light blue, and orange) are initialized (top row). These are passed through z1...z100. At each stage, the variable is updated by reducing its value by √(1-β) and adding noise with zero mean and variance β (Equation 1). This process causes each of the points to "move" towards a distribution centered around zero.

**(b)** Probability distributions of variables at selected time stages z1, z41, z81. As the diffusion process progresses, the distributions expand, their means approach zero, and they increasingly overlap. The trend is convergence towards a standard normal distribution.

**To train the decoder to reverse the process, multiple examples of Zt at time t for the same sample X can be used. Generating the variable Z sequentially using Equation 2 is time-consuming when T is large. However, there is a closed-form expression for q(Zt|X), allowing direct sampling from Zt given the initial data point X without calculating intermediate variables Z1,…, ZT-1.**

**To see this, we will substitute t=2 into Equation 2 and obtain Equation 3:**

**Then place Equation 1 into Equation 3 to get Equation 4:**

**The noise scheduler can be written differently, yielding Equation 4.1:**

**We can simplify further to obtain Equation 4.2:**

The last two terms are independently sampled from normal distributions with zero mean and variance (1−β2−(1−β2) (1−β1)) and β2, respectively. The mean of this sum is zero, and its variance is the sum of the variances of the components. Additionally, given that ϵt is a random noise in every state and since the sum of normal distributions is normal with variance equal to the sim of the variances, we get Equation 5.

If we continue this process for t=3, and so on, we can add a variable α defined in Equation 6:

Thus, obtaining Equation 7:

Equation 7 is important because it shows that for any initial data point X, the variable Zt is normally distributed with known mean and variance. As a result, if we do not care about the history of the evolution through intermediate variables Z1,…, Zt-1, it is easy to sample from q(Zt|X). So even if T is very large, under the right parameter settings and consistent constraints such as variance, mean, and variable α, we can calculate the noise for each t separately without always sampling є but directly reaching Zt.

**Decoding:** When training in the diffusion process, we learn the reverse process. This means we learn a series of probabilistic mappings back from the variable Zt to Zt−1, from Zt−1 to Zt−2, and so on until we reach the data X. The true reverse distributions q(Zt−1|Zt) of the diffusion process depends on the data distribution Pr(X). The model needs to approximate these distributions as normal distributions, resulting in Equations 8, 9, and 10 for the early, middle, and final decoding stages, respectively:

where ft [Ztϕt] is a neural network computing the mean of the normal distribution in the estimated mapping from Zt to the previous variable Zt-1. The terms {σ t} are predetermined. If the hyperparameters βt in the diffusion process are close to zero (and the number of time steps T is large), then this normal approximation will be reasonable. We generate new samples from Pr(X) by sampling.

* Φ = the neural network parameters, including learned weights guiding the denoising process.
* f = a function learned by a neural network predicting the mean of the next variable.
* I = identity matrix indicating that noise is independently added to each dimension.
* σ = standard deviation of the noise at each stage of the reverse process.
  1. **Algorithm**
     1. **Forward process (encoder)**

|  |  |
| --- | --- |
| **Algorithm 1** | Forward Process in diffusion model |
| **Input:** | Data X |
| **Output:** | Sequence of noisy data {Zt} |
| for t = 1 to T do: |  |
| є ~ Norm [0, I] | // sample noise |
| Zt = sqrt(1-βt) \* Zt-1 + sqrt(βt) \*є | // add noise to data |
| end for | |

The forward process of a diffusion model can be written in pseudo code like this:

During this process, Gaussian noise is added to the data X for T steps, creating a sequence of increasingly noisy data representations {Zt}. The result for each individual data point X can be seen in Figure 5. However, this type of algorithm might have high computational costs and inefficiencies when processing high-dimensional data, as these costs scale significantly with the data's dimensionality. A solution to this can be found in Latent Diffusion Models (LDM).

**Latent Layers:** Latent layers in an LDM transform the input data through a series of intermediate states, effectively compressing it into a lower-dimensional space. This process is almost completely the same as the non-latent layers except for:

1. The encoder takes the input data X and maps it to the latent space. It is then passed through multiple layers to produce a series of latent variables Z1,Z2,…, Zt.
2. At each layer, Gaussian noise is added to the latent variables, gradually destroying the structure of the data. And finally, after T steps, the final latent variable ZT follows a standard normal distribution.

The forward process for a latent diffusion model can be written in pseudo code like this:

|  |  |
| --- | --- |
| **Algorithm 2** | Forward Process in latent diffusion model |
| **Input:** | Data X |
| **Output:** | Sequence of latent representations {Zt} |
| Z0 = Encoder(X) | // map X to the latent representation |
| for t = 1 to T do: |  |
| є ~ Norm [0, I] | // sample noise |
| Zt = sqrt(1-βt) \* Zt-1 + sqrt(βt) \*є | // add noise to latent rep representation |
| end for | |

In this approach, the data X is first mapped to a lower-dimensional latent representation Z0 using an encoder. Then, similar to the standard diffusion model, Gaussian noise is added to the latent representation over several steps, generating a sequence of noisy latent representations {Zt​}. This reduces the computational burden by working in a lower-dimensional latent space instead of the high-dimensional original data space.

* + 1. **Reverse process (decoder)**

The reverse process or decoding process is the core mechanism of the Diffusion Model (DM). This process is where the model iteratively removes noise added during the forward process, effectively reversing the diffusion steps to recover the original data.

Each step involves using a neural network to predict and subtract the noise, gradually converging on the original data distribution.

To do so, we can use a neural network as we seen in Equation 10 where we had ft[Zt,ϕt] for a neural network that computes the mean of the normal distribution in the estimated mapping from Zt to the previous latent variable Zt-1. For that type of neural network, we can use a U-Net model.

* **U-Net:**

U-Nets are a type of convolutional neural network (CNN) introduced in a 2015 paper and initially designed for biomedical image segmentation [10].

The architecture's name derives from its U-shaped design, where the encoding and decoding paths are symmetrical. The contracting path of the U-Net follows the typical architecture of a convolutional network.

* **Convolution:**

Convolution is an operation in CNN where a filter or kernel slides over the input data to produce a feature map. The filter is a small matrix, and the convolution operation involves element-wise multiplication followed by summation.

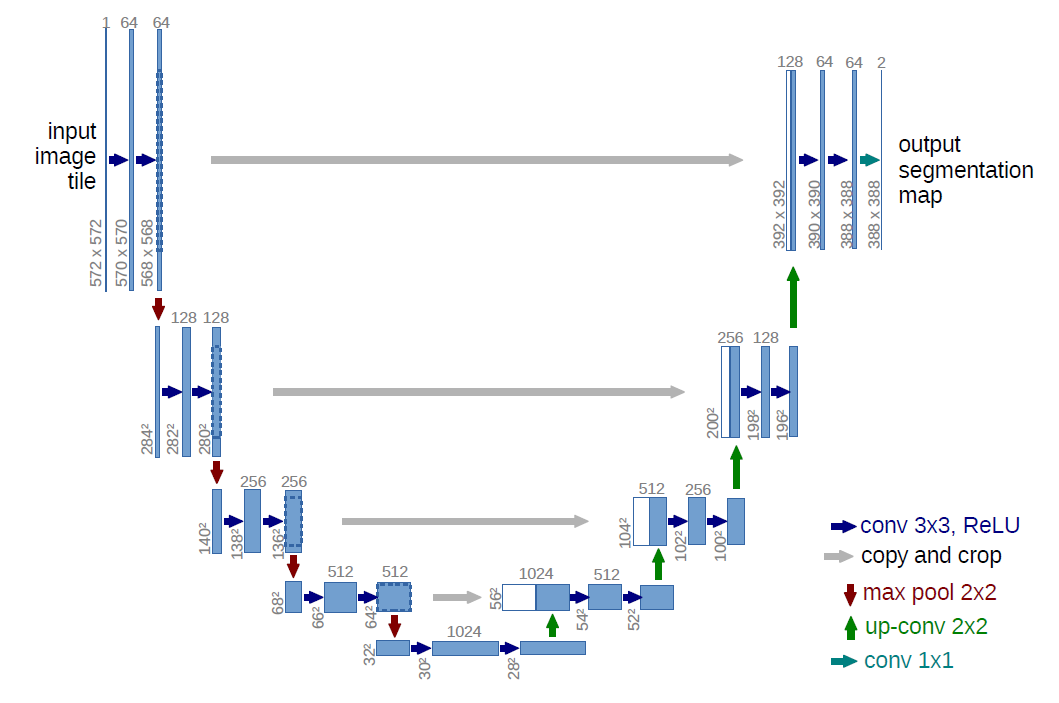
This process allows the network to learn spatial hierarchies of features from the input image. In U-Nets, each layer in the contracting path consists of the repeated application of two 3x3 convolutions, which are unpadded. The convolutional filters capture local patterns, such as edges or textures, which are crucial for understanding the structure of the input image. Each convolution is followed by a rectified linear unit (ReLU), which introduces non-linearity into the model, allowing it to learn more complex representations.

Following the convolutions, the network employs a 2x2 max pooling operation with a stride of 2 for downsampling.

* **Pooling:**

Pooling is another essential operation in CNNs that reduces the spatial dimensions of the feature maps while retaining the most critical information. The pooling operation selects a statistic of each region of the feature map, e.g. average, maximum, effectively reducing the size of the feature map. This reduction helps in achieving translation invariance, making the network more robust to variations in the input data. In the U-Net architecture, there is a usage of max pooling 2x2 operation which selects the maximum value from each 2x2 region of the feature map, effectively reducing the size of the feature map by a factor of two. This max pooling operation is used to progressively reduce the spatial dimensions of the feature maps in the contracting path, allowing the network to capture high-level features over larger regions of the input image.

The U-net network architecture is illustrated in Figure 7. It consists of a contracting path (left side) and an expansive path (right side). The contracting path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions, each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step, the spatial dimensions of the input are reduced while the number of channels (feature maps or depth), which represent different aspects of the input, is doubled.



**Figure 7 (adapted from [10], Fig. 1):** U-Net layers architecture and each operation applied on each layer.

In the **early layers**, low-level features such as edges, textures, and simple patterns are captured. These layers are crucial in the initial stages of image analysis, focusing on fine details. As the input image passes through these layers, its spatial dimensions are reduced while the number of channels increases, allowing the network to capture detailed local patterns.

The **bottleneck layers** arelocated at the deepest part of the U-Net, where the information is compressed into a highly abstract representation. This compression significantly reduces the spatial dimensions while further increasing the number of channels. The bottleneck layers handle the most condensed form of the data, emphasizing high-level, abstract features over fine details.

Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU.

In the **late layers**, the compressed information from the bottleneck layers is integrated and decoded, refining the details and reconstructing the image to its original resolution. These layers handle more complex, high-level features, ensuring that the final output is both accurate and detailed.

At the final layer, a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In total, the network has 23 convolutional layers, effectively balancing the preservation of detailed spatial information and broader contextual understanding [10].

Thus, the reverse process algorithm for a latent diffusion model can be written in as:

|  |  |
| --- | --- |
| **Algorithm 3** | Reverse Process in latent diffusion model |
| **Input:** | Sequence of latent representations {Zt} |
| **Output:** | Reconstructed data X |
| for t = T to 1 do: |  |
| 𝜖θ = Model(Zt, t) | //predict noise using a neural network |
| Zt-1 = (Zt - sqrt(βt) \* 𝜖θ) / sqrt(1 - βt) | // denoise step |
| end for |  |
| X = Decoder(Z0) | // map Z0 to the original data space |
| Return X | // returns the reconstructed data |

In this methodology, we begin with the final noisy representation ZT. We employ a neural network model to predict the noise that must be subtracted from ZT to obtain ZT-1. This process is iteratively continued from ZT-1 to ZT-2, and so forth, until we reach Z0. Once Z0 is obtained, it is decoded back to a higher-dimensional space than the latent representation, resulting in X, which is our reconstructed data.

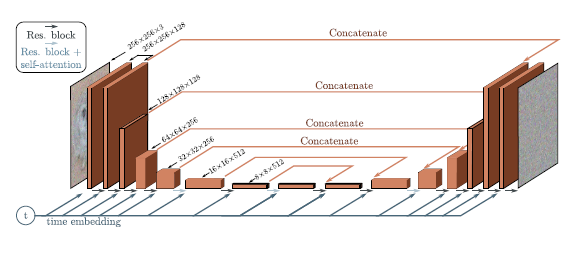
* 1. **Application for Generation**

The U-Net model is widely regarded as highly effective due to its unique architecture, which allows it to integrate both high-resolution and low-resolution details throughout the network. Traditional CNNs reduce dimensions from initial features to high-level features, transitioning from shape and texture to semantic meaning. U-Net preserves both micro and macro details with concatenation. This approach is similar to the brain's natural architecture. For example, the human visual system processes detailed information (such as edges and textures) in the primary visual cortex while simultaneously integrating broader contextual information (such as the overall scene) in higher-order visual areas. This retention of detailed spatial information while capturing broader context makes U-Net powerful for tasks needing both fine details and overall context [11].

**Concatenate -** concatenation occurs in the upsampling pathway, where feature maps from the encoder’s contracting path are concatenated with the corresponding feature maps from the decoder’s expansive path. This helps to preserve the spatial information that may be lost during the downsampling, thereby providing detailed context [10, 11].

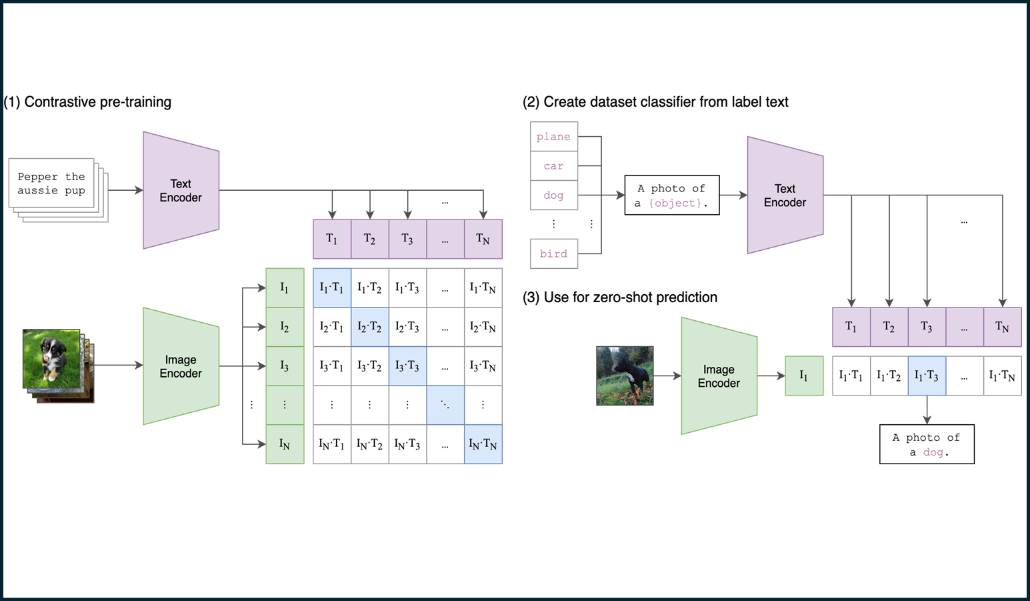
**Time embedding -** Time embedding is essential for deciphering noise in non-transformer networks. The network must ascertain the specific time point at which the input was received corresponds to the noise level as a function of time t. A vector representation of the time variable t is generated and propagated through the entire network. ensuring each layer incorporates and learns this temporal information. [8, 10] An illustration of such network can be seen in Figure 8 [10].

**Figure 8 (adapted from [9], Fig. 18.9):** The figure illustrates the U-Net architecture enhanced with time embedding. The architecture consists of two main pathways: the contracting path on the left and the expansive path on the right.



**Clip -** CLIP (Contrastive Language-Image Pre-training) is a model developed by OpenAI that encodes both text and images into a shared latent space. This allows the model to measure the semantic distance between textual descriptions and visual content. The core idea behind CLIP is to leverage a large dataset of images paired with their corresponding textual descriptions to train a model that can understand and relate both modalities [1].

The CLIP model consists of two main components: an image encoder and a text encoder. The image encoder processes the input image and transforms it into a fixed-length feature vector. The text encoder converts the input text into a fixed-length feature vector as well. During training, CLIP learns to maximize the similarity between the feature vectors of matching image-text pairs and minimize the similarity between mismatching pairs [12].



**Figure 9(adapted from [12], Fig. 1):** This figure illustrates the architecture of the CLIP model, highlighting the interaction between the text encoder and image encoder through a contrastive learning framework.

In a Unet, the semantic representation of an image (a vector representing the text describing the image) can be propagated through each layer of the network, similar to how time embedding is done - A clip variable c is propagated throughout the network according to the obtained representation. This propagation ensures that the model learns to incorporate and understand the semantic information at each layer, enhancing its ability to relate text and images effectively [2].

1. **Visual Information Reconstruction in Neural Networks**

The early attempts of visual information reconstruction from neural signals relied on simple, linear models that used basic features of neural activity, such as the firing rates of specific neurons, to infer the visual stimuli being perceived. These models were limited in resolution and accuracy [7].

A significant advancement in this field was made by Naselaris et al. (2009), who utilized functional magnetic resonance imaging (fMRI) data to reconstruct images that subjects were viewing. By correlating voxel activity with specific image features, the researchers were able to generate rudimentary reconstructions of the viewed images. Despite the low resolution and high noise levels, this study laid crucial groundwork for future research in neural image reconstruction [13].

Further advancements were demonstrated by Lin et al. (2023), who developed a method for reconstructing complex images from brain activities by leveraging advanced machine learning algorithms. Their approach involved mapping fMRI data into embeddings in the CLIP space, which were then used to condition a model to generate images. The results showcased high-resolution and more accurate image reconstructions compared to previous methods, demonstrating the potential of integrating computational techniques with neuroimaging data to enhance the fidelity of visual reconstructions [14].

In their 2023 study, Takagi and Nishimoto introduced a novel approach in the exploration of image reconstruction using stable diffusion. Their research presents a method for reconstructing images from human brain activity, specifically utilizing fMRI. The authors delve into the internal mechanisms of LDMs by examining the relationships between key components—such as the latent image vector (Z), conditioning inputs (C), and elements of the denoising U-Net—and various brain functions. Their findings demonstrate that this method can generate high-resolution images with remarkable fidelity without the need for additional training or fine-tuning of complex models. Furthermore, they provide a quantitative interpretation of the LDM components from a neuroscientific perspective, offering new insights into the intersection of machine learning and neuroscience [2].

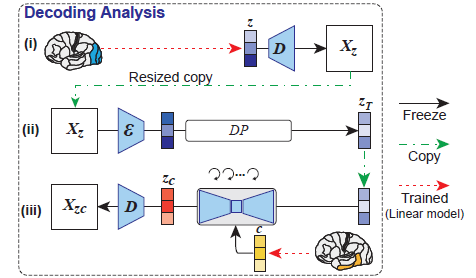
The researchers utilized a dataset which comprises brain activity recordings from subjects viewing 10,000 images over multiple sessions using a fMRI scanner. Data from four subjects were analyzed, with each subject viewing three repetitions of these images. That means that each scan corresponds to specific images viewed by the subjects, allowing researchers to map the neural responses to these visual inputs. The fMRI data is measured in voxels, the images that were shown to the subjects were resized to 425x425 pixels, and the dataset included 27,750 trials per subject, which were divided into training and test sets. Preprocessed scans with a were used to explore the link between visual stimuli and brain activity, aiding in the reconstruction of visual images from the brain data.

The authors employed a LDM and trained it on the dataset. Actually, the only training required in this method is to construct linear models that map fMRI signals to each LDM component, and no more training or fine-tuning of deep-learning models is needed. They define Z as the latent representation of the original image compressed by the autoencoder, c as the latent representation of texts, and Zc as the generated latent representation of Z modified by the model with c.

The reconstruction process of this study had three key steps:

1. **Predicting Latent Representation**: The latent representation Z of the presented image X was predicted from fMRI signals in the early visual cortex. This z was then processed by the autoencoder's decoder to produce a coarse image Xz​ at 320x320 pixels, which was resized to 512x512 pixels. So, first they predicted a latent representation x of the presented image X from fMRI signals within early visual cortex.
2. **Noise Addition**: The coarse image Xz​ was further processed by the autoencoder's encoder and underwent a diffusion process to add noise.
3. **Decoding Text Representations**: Latent text representations cc was decoded from fMRI signals in the higher visual cortex. The noise-added latent representations zT​ and c were input into the denoising U-Net to generate Zc​, which was then decoded by the autoencoder to produce the final reconstructed image Xzc​ at 512x512 pixels.

In order to build models from fMRI into components of LDM, the authors used L2-regularized linear regression - a statistical modeling technique which enhances linear regression by adding a penalty term to the loss function to prevent overfitting and improve generalization - and Weights were estimated from training data. As control analyses, they also generated images using only z or c. To generate these control images, they simply omitted c or z from step (iii) above, respectively.



**Figure 10(adapted from [2], Fig. 2**) The figure illustrates the reconstruction process of the study in three key steps:

***i. Prediction of Latent Representation Z:*** *fMRI signals predict Z of image X. Z is decoded to create coarse image Xz.*

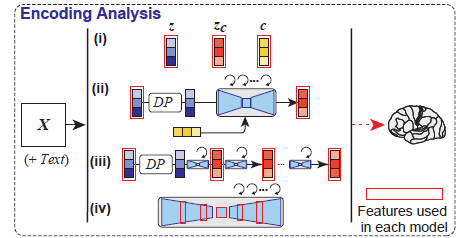
***ii. Noise Addition through Diffusion:*** *Xz is encoded, diffused, and transformed into noise added ZT.*

***iii. Decoding of Latent Text Representations c:*** *Text representations c are decoded from fMRI. ZT and c are inputs to a U-Net, producing Zc, which is decoded to the final image Xzc.*

The accuracy of image reconstruction was evaluated objectively using Perceptual Similarity Metrics (PSMs) and subjectively by human raters. PSMs assess the perceptual similarity between two images based on human visual perception, rather than just pixel-by-pixel comparison. These metrics capture elements like texture, structure, and semantic content using models (AlexNet, etc.) that process images in a way that mimics human visual perception, extracting features from different layers of the network. The evaluation involved determining whether the original test images could be identified from the generated images. As a similarity metric for PSMs, early, middle, and late layers of CLIP were used. Early layers usually capture low-level features such as edges and textures. Middle layers capture more complex patterns and structures, and Late layers capture high-level semantic information, such as object categories and scene context. In conclusion, a two-way identification experiment examined whether the image reconstructed from fMRI was more similar to the corresponding original image than to a randomly selected reconstructed image.

The researchers aimed to elucidate the internal mechanisms of Latent Diffusion Models (LDMs) by correlating them with human brain activity through whole-brain voxel-wise encoding models. Their study encompassed four primary settings:

1. **Independent Predictions**: They predicted voxel activity independently from three LDM latent representations (z, c and Zc).
2. **Combining Predictions**: Despite Zc and z generating distinct images, they resulted in similar brain activity patterns. By integrating these predictions into a single model and mapping unique variance, the authors-controlled noise levels and interpreted the image-to-image processing.
3. **Denoising Process**: To investigate the internal dynamics of the denoising process, Zc was extracted at early, middle, and late stages. Combined models with z were then constructed to map their unique variance.
4. **U-Net Layers**: Features were extracted from various U-Net layers during the denoising process. Encoding models were developed for different denoising steps, identifying the most accurate layer for each voxel and step.



**Figure 13(adapted from [2], Fig. 2**) Schematic of encoding analysis with model encoding options to predict fMRI signals from different components of LDM, including Z, C and Zc,

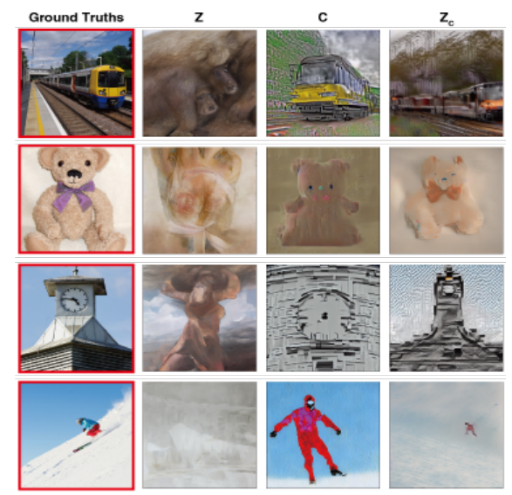
Overall, this methodology aimed to quantify and interpret the image reconstruction process of LDMs, thereby bridging the gap between brain activity and the operations of complex neural networks.

For evaluation of the model's ability to predict fMRI signals, Pearson’s correlation coefficients were used to compare predicted and actual fMRI signals. They checked statistical significance(one-sided) by comparing these correlations to those of two independent Gaussian random vectors of the same length.

By comparing the observed correlations to this null distribution, the researchers can determine if the observed correlations are significantly greater than what would be expected by random chance. The significance threshold was set at P<0.05, meaning there is less than a 5% chance that the observed correlations are due to random variation and corrected for multiple comparisons using the False Discovery Rate method (FDR), which controls the expected proportion of incorrectly rejected null hypotheses (false positives), ensuring that the findings remain robust even when multiple statistical tests are conducted.

* 1. **Results**

We can see in figure 12 the results of one of the subjects. We can observe that the images reconstructed using only Z were visually somehow consistent with the original images – the basic structure and layout of the images were preserved. However these reconstructions failed to capture the semantic content, such as specific objects and details that make the image meaningful – what makes a train a train or a teddy bear a teddy bear.



**Figure 12(adapted from [2], Fig. 3)** This figure shows the presented (red box) and reconstructed images for a single subject using Z, c, and Zc**.**

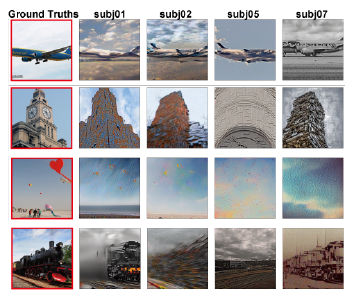
On the other hand, the images reconstructed using only C generated images with high semantic fidelity but were visually less precise, leading to distortions or alternations in the appearance.

For example, we can see that the model was able to reconstruct the train quite well but lost its orientation and the relative direction from the original image (Ground Truths) images reconstructed using When both Z and C were combined to form Zc, the reconstructed images managed to achieve a balance, producing high-resolution images that were both visually consistent and semantically accurate.

This combination leveraged the structural information from Z and the detailed semantic information from C to create reconstructions that closely resembled the original images in both form and content. We will see more of that Z, C Zc dynamic shortly.

In figure 13 we can observe several reconstructed images from four different subjects that were generated using Zc. These reconstructions show that the approach is generally effective across different individuals. However, there are some discrepancies visible in certain images.

**Figure 13(adapted from [2], Fig. 4)** Reconstructed images from different subject for the same image generated using Zc.



These discrepancies could arise from various factors. For instance, differences in perceived experience across subjects might influence how their brain activity correlates with the images, leading to variations in the reconstructed outputs. Additionally, potential failures in the model's reconstruction process could introduce inconsistencies.

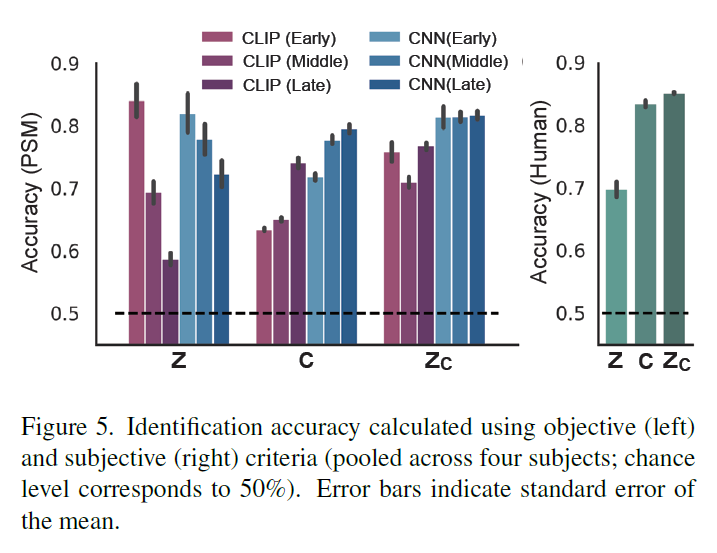
Another important factor is the difference in data quality among subjects, which can significantly impact the accuracy of the reconstructions. Subjects with higher quality fMRI data tend to have more accurate reconstructions, while those with lower quality data might exhibit more discrepancies. Despite these challenges, the use of Zc generally provided a robust method for generating high-resolution, semantically faithful reconstructions from brain activity.

Going back to evaluating Z, C and Zc, Figure 14 plots results for the quantitative evaluation. In the objective evaluation (PSM), images reconstructed using Zc generally exhibited higher accuracy values across various metrics than images reconstructed using only Z or C.

When only z was employed, the accuracy values were notably high for PSMs derived from early layers of CLIP and CNN. This indicates as we saw in figure 12, that the structural and textural features captured by these early layers were well-preserved in the reconstructions using Z. Conversely, when only c was used, accuracy values

were higher for PSMs derived from late layers which more prominently captures the semantic content and high-level features.

**Figure 14(adapted from [2], Fig. 5)** Identification accuracy calculated using objective (CNN, CLIP) and subjective (human raters) criteria.



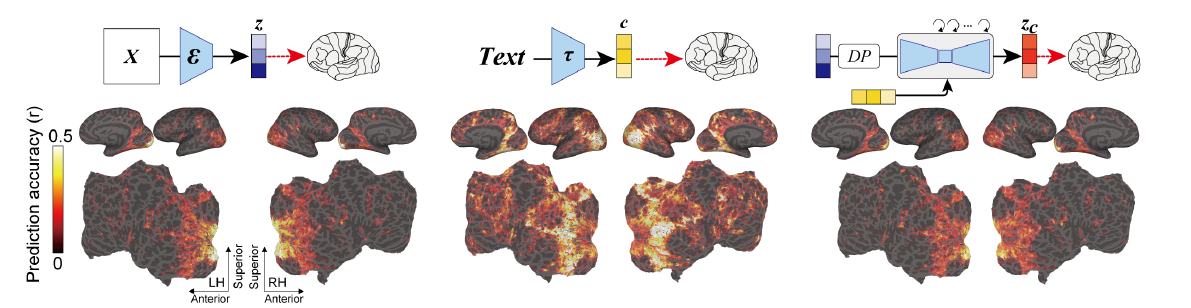
In the subjective evaluation (Human), accuracy values for images reconstructed using C were higher than those using Z. However, the highest accuracy was achieved with images reconstructed using Zc, outperforming the other two methods. This combination effectively integrates both the low-level visual details and high-level semantic content.

These findings collectively suggest that this method is capable of capturing not only the low-level visual appearance but also the high-level semantic content of the original stimuli. This dual capability underscores the robustness of this approach in accurately reconstructing images from brain activity.

Despite all three components showing high predictive performance in the posterior part of the brain, specifically the visual cortex, they exhibited notable differences. These differences were studied in different "location" or fields in this study: Latent representation, Noise level, Diffusion stage, U-net layers.

* **Latent representation**

Differences in the Prediction accuracy (r) of latent representations can be seen in Figure 15. Specifically, Z demonstrated high predictive performance in the posterior visual cortex, particularly in the early visual cortex. It also showed significant predictive values in the anterior visual cortex, specifically the higher visual cortex, but smaller values in other regions. On the other hand, c produced the highest predictive performance in the higher visual cortex. The model also showed high predictive performance across a wide range of the cortex.



**Figure 15(adapted from [2], Fig. 6)** Prediction accuracy (r) of encoding models for types of latent representations in the LDM: z, c, and Zc. The brain maps highlight regions with varying predictive accuracy, with z and Zc showing high performance in early visual cortex and c excelling in higher visual cortex.

Zc, bearing a representation very similar to z, showed high predictive performance for the early visual cortex. Although this is somewhat expected given their intrinsic similarity, it is nonetheless intriguing because these representations correspond to visually distinct generated images.

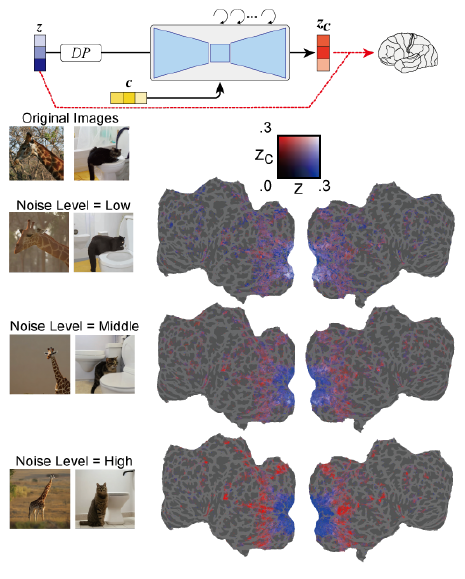
* **Noise level**

Despite the previous results showing similar prediction accuracy maps for Z and Zc, they do not reveal how much unique variance each feature explains as a function of different noise levels. This means that while Z and Zc seem to perform similarly overall, it is difficult to understand the individual contributions of Z and Zc at various levels of noise. To address this problem, encoding models that simultaneously combined both Z and Zc into a single model and studied the unique contribution of each feature were constructed. Combining Z and Zc in a single model allows to see how each one uniquely affects the predictions. This approach helps in isolating the specific roles that each representation plays in the overall prediction and can be seen in Figure 16

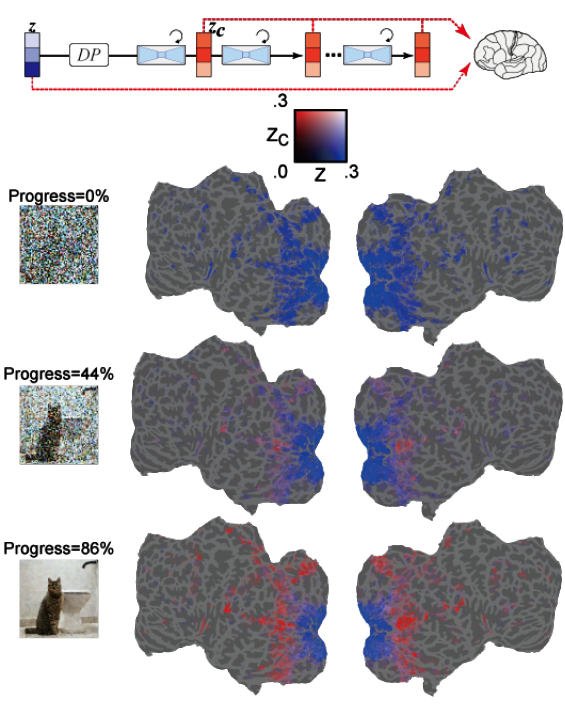
The noise level added to Z to create Zc was varied, Varying the noise levels helps to understand the robustness of each feature and allows to see how well Z and Zc perform under different conditions, providing insights into their strengths and weaknesses.

Eventually when a small amount of noise was added, Z predicted voxel activity better than Zc across the cortex. Interestingly, as the noise level was increased, Zc predicted voxel activity within the higher visual cortex better than Z, indicating that the semantic content of the image was progressively emphasized, meaning that Z performs better with low noise, while Zc excels in the higher visual cortex as noise levels increase. This indicates that as more noise is added, the model shifts from focusing on detailed visual features to capturing the overall meaning or context of the image. This shift highlights the adaptability of Zc in emphasizing semantic content over visual details under noisy conditions.

This result is intriguing because, without analyses like this, we can only observe the randomly generated images and cannot examine how text-conditioned image-to-image processes balance between semantic content and original visual appearance.



**Figure 16(adapted from [2], Fig. 7)** Figure 16 shows unique variance accounted for by Zc compared with Z across varying noise levels.



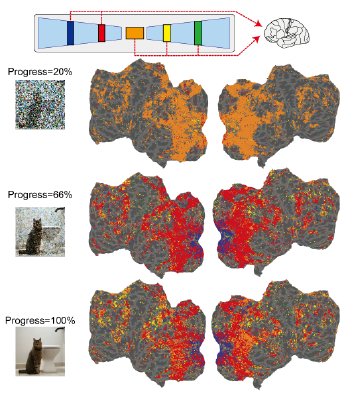
**Figure 17(adapted from [2], Fig. 8)** The denoising stages of LDM displayed in voxel activity predictions at different stages of noise reduction.

* **Diffusion stage**

Figure 17 shows the process by which Latent Diffusion Models (LDMs) refine and generate images from noise in different diffusion stages. During the early stages of the denoising process (0% progress), Z signals dominated the prediction of fMRI signals, indicating that detailed visual information is crucial at this stage. As the denoising progresses to the middle stage (44% progress), Zc begins to predict activity within the higher visual cortex much better than Z, signifying that the bulk of the semantic content emerges at this point. This transition highlights how the LDM shifts its focus from low-level visual details to high-level semantic content as noise is reduced, refining the image progressively and mirroring the hierarchical nature of human visual processing.

* **U-net layers**

During the early phase of the denoising process, the bottleneck layer of U-Net shows the highest prediction performance across the cortex. As denoising continues, the early layer of U-Net predicts activity in the early visual cortex, while the bottleneck layer shifts to having superior predictive power in the higher visual cortex. These findings suggest that, at the start of the reverse diffusion process, image information is compressed in the bottleneck layer. As denoising advances, a functional differentiation among U-Net layers appears in the visual cortex: the first layer typically represents fine-scale details in early visual areas, while the bottleneck layer corresponds to higher-order information in more ventral, semantic areas.



**Figure 18(adapted from [2], Fig. 9)** results of encoding models for different steps of the denoising process (early, middle, late), and

*for the different layers of U-Net*

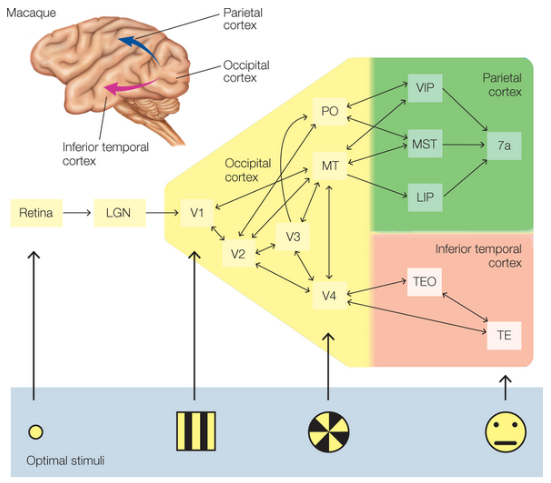
* 1. **Biological interpretation**

Biological interpretation is a nuanced endeavor. Can we extrapolate experimental results to our neuronal or cognitive systems? Can we assert that the brain processes visual information similarly to the cognitive models we have studied, especially when these models yield impressive and high-resolution results? Most likely not and drawing such parallels without careful consideration could be a serious error. Although neural network models are inspired by brain function and strive to mimic its processes, they do not perfectly replicate the brain's operations. These models, while advanced and informative, serve as approximations rather than exact replicas of neuronal activity [15].

Hence, caution is warranted when interpreting their results as direct analogs of brain function. Experimental results from these models should be viewed as illustrative insights rather than definitive representations of cognitive processes. Furthermore, the importance of interpretability in these models cannot be overstated. It allows researchers to understand how conclusions are derived, ensuring that the insights gained are meaningful and can be validated against biological systems, rather than being treated as black-box predictions [15, 16].

Considering this, certain fields, such as computer vision, endeavor to develop systems that perceive and recognize the world in a manner similar to humans. Advances in brain activity measurement and deep neural network design facilitate comparisons between latent representations in biological brains and the architectural characteristics of artificial networks. To assess this, we can investigate patterns and differences in the behaviors of artificial models and biological brains.

The findings from our study underscore that the model demonstrates a hierarchical processing architecture, where features are processed sequentially from simpler to more complex forms. This does somehow mirror the processing in the visual system, where the complexity of the optimal stimulus increases progressively along the pathway: retinal and lateral geniculate nucleus (LGN) cells predominantly respond to simple light spots, whereas primary visual cortex (V1) cells exhibit sensitivity to edges. In higher cortical areas, such as V4, the cells process more intricate stimuli including shapes and faces [17].



**Figure 19(adapted from [17], Fig. 5.25)** This figure shows the progression of visual information processing within the macaque brain. As visual information ascends through the visual processing hierarchy the complexity of the stimuli that elicit the strongest neuronal responses increases.

This clear alignment with a bottom-up processing model, where complexity escalates as information moves from lower to higher processing stages, highlights the potential of these models in replicating and understanding visual perception in brain activity. However, it's important to note that this does not provide the full picture, as top-down processes also play a critical role in visual perception. For instance, expectations and prior knowledge can influence how visual information is interpreted, such as when recognizing ambiguous images or filling in visual gap [18].

Our findings illustrate the potential of Latent Diffusion Models to reconstruct detailed visual perceptions from brain activity, mimicking some aspects of hierarchical visual processing observed in natural neural systems. While these results are promising, they should be interpreted with caution; the analogy between computational models and biological processes is not direct. Future research could focus on refining these models to improve their accuracy and interpretability.

1. **Discussion**

Comparing these results with previous studies is challenging due to differences in datasets. There is one study though, that used NSD for visual reconstruction [19]. It is difficult to draw a direct comparison with this study, However, this prior study relied on extensive model training and feature engineering with many more hyper-parameters, including the necessity to train complex generative models, fine-tuning' data augmentation, and arbitrary thresholding of features.

In a study by Nishimoto et al. (2011), a model was developed to reconstruct natural movies from evoked BOLD signals. When subjects watch new movies, their brain activity is recorded using fMRI. The model decodes this activity by identifying which visual stimuli (movie frames) most likely correspond to the recorded patterns. This method yields a low-resolution video approximating the subject's visual experience. [20].

Future advancements in this field may leverage findings from such studies to enhance the resolution of video reconstructions. This implies that reconstructions could extend beyond simple images or videos, potentially reproducing the continuity that characterizes the realm of thoughts in high resolution. Additionally, an OpenAI model named Sora has recently been introduced. This model is based on similar principles to their DALL-E image reconstruction model, incorporating elements from Latent Diffusion Models (LDMs) discussed earlier. The application of this technology may facilitate the prediction of brain states and mental states, enabling the reconstruction of specific videos to which the subject was exposed [21].

The successful reconstruction of images from brain data highlights a fascinating possibility: thoughts may have a structured, decipherable format. This advancement in neural imaging and computational modeling suggests that our internal visual experiences can be translated back into images, hinting at a methodical way to interpret brain activity. As techniques like Latent Diffusion Models become more refined, they not only illuminate the process of how the brain encodes visual information but also pose profound questions about the extent to which we can decode and understand the broader landscape of human thoughts.

The capacity to reconstruct images from brain activity invites a philosophical inquiry into the nature of thoughts themselves. If we can decode visual information from neural patterns, it may imply that thoughts are not just ephemeral and subjective experiences but have a definite structure that can be systematically accessed and interpreted. This idea challenges traditional views of the mind as an impenetrable fortress and suggests a model where thoughts could be quantified and potentially manipulated. Such a prospect raises ethical considerations about privacy and the integrity of personal mental space, while also opening up possibilities for therapeutic interventions in neurological and psychological disorders. The question remains, however, whether this technological capability truly captures the essence of thought or merely its shadows, an echo of our mental processes rather than the substance itself.

In summary, computational models function by processing input data through a sequence of computational steps dictated by a specified algorithm, culminating in an output. This output, or the model's "behavior," is then juxtaposed with empirical data derived from methodological research. Such comparisons facilitate an assessment of the model's characteristics and its predictive capacity concerning human behavior. The primary objective of these models is to mathematically represent and elucidate human behavior. However, this does not imply that the brain processes an image, reconstructs memories, or generates thoughts in a manner akin to the steady diffusion model; rather, insights can be gleaned from the model's behavior.

We examine an innovative visual reconstruction technique using Latent Diffusion Models (LDMs), which demonstrates that this method can reconstruct high-resolution images with considerable semantic accuracy from human brain activity. Contrary to prior image reconstruction studies, this method eschews the need for the training or fine-tuning of intricate deep-learning models, relying solely on straightforward linear mappings from functional magnetic resonance imaging (fMRI) data to latent representations within LDMs.

We reviewed a quantitative interpretation for the internal components of the LDM by building encoding models. For example, we observed the emergence of semantic

content throughout the inverse diffusion process, we also observed a layer-wise characterization of U-Net, and with a quantitative interpretation of image-to-image transformations with different noise levels. This study is the first to provide a quantitative interpretation from a biological perspective. Eventually, when z, Zc, and c behave differently across the stages of the model, it reveals their unique roles. Z captures the structural aspects of images, c encodes semantic details, and Zc combines these to produce precise reconstructions. This differentiation aids in understanding how the brain might process visual information in stages, reflecting a progression from basic perception to complex interpretation, similar to how the model operates. As the model progresses, Zc becomes more predictive of activity in the higher visual cortex, showing its increasing role in integrating semantic content. This transition from Z to Zc highlights the model's ability to evolve from simple to complex information, mirroring the brain's method of gradually refining visual inputs to extract deeper

1. **References**
2. Aditya R., Prafulla D., Alex N., Casey C., Mark C. (2022) Hierarchical Text-Conditional Image Generation with CLIP Latents
3. Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity.
4. (2022). פסיכולוגיה קוגניטיבית: נושאים נבחרים בחקר התפיסה [Cognitive Psychology: Selected Topics in Perception].
5. Carlson, N. R., & Birkett, M. A. (2018). The Physiology of Behavior (12th ed., Vol. 1, Hebrew trans.). The Open University of Israel. (Orig. pub. 2017 by Pearson Education, Inc.).
6. Belliveau, J. W., Kennedy, D. N., McKinstry, R. C., Buchbinder, B. R., Weisskoff, R. M., Cohen, M. S., . . . Rosen, B. R. (1991). Functional mapping of the human visual cortex by magnetic resonance imaging. Science, 254(5032), 716–719.
7. Li, Q., Del Ferraro, G., Pasquini, L., Peck, K. K., Makse, H. A., & Holodny, A. I. (2020). Core language brain network for fMRI language task used in clinical applications. Netw. Neurosci. 4(1), 134-154. <https://doi.org/10.1162/netn_a_00112>
8. Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. Nature, 452(7185), 352–355.
9. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics.
10. Prince, S. J. D. (2023). Understanding Deep Learning. The MIT Press.
11. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), (pp. 234-241).
12. Xu, Q., Ma, Z., Na, H. E., & Duan, W. (2023). DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation. Computers in Biology and Medicine.
13. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.
14. Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. Neuron, 63(6), 902-915.
15. Lin, S., Sprague, T., & Singh, A. K. (2023). Mind Reader: Reconstructing complex images from brain activities. UC Santa Barbara
16. Schwessinger, R., Kuhn, T., Luu, M., Woodhouse, M. R., Gottlieb, D., Dinh, V., & Thorsen, T. (2022). Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biology, 23(1).
17. Hanczar, B., Zehraoui, F., Issa, T., & Arles, M. (2020). Biological interpretation of deep neural network for phenotype prediction based on gene expression. BMC Bioinformatics, 21, 501.
18. Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2019). Cognitive Neuroscience: The Biology of the Mind (5th ed).
19. Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. Nature Reviews Neuroscience, 14(5), 350-363.
20. Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. Advances in Neural Information Processing Systems, 9 2022.
21. Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology, 21(19), 1641-1646.
22. OpenAI. (2024). Sora: Text-to-Video AI Model.