Applications of Diffusion Models

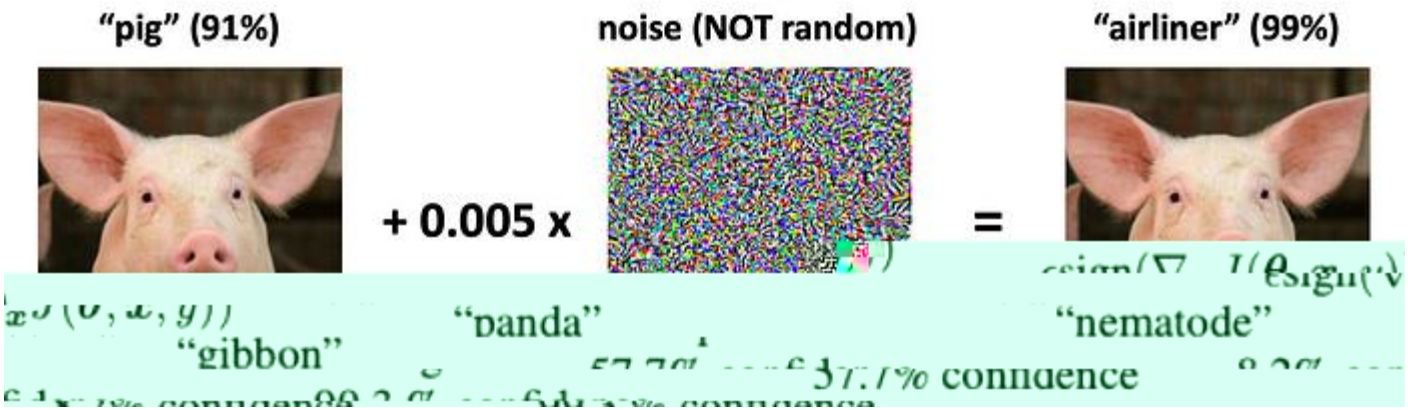
Jian Li

IIIS, Tsinghua

ATCS 2023

Adversarial Robustness and Score-based generative model

A Brief Introduction to Adversarial Robustness



Adversarial Purification:
Another approach for the adversarial defense is to purify attacked images before feeding them to classifiers

Adversarial Purification

‡ Learn a purification model whose goal is to remove any existing adversarial noise from potentially attacked images into clean images so that they could be correctly classified when fed to the classifier.

‡ The purification model is usually trained independently of the classifier

‡ The most common way is for adversarial purification to learn a generative model:

Generate clean images from attacked images.

High level idea

- Adversarial purification can be understood as a denoising procedure
- Learning scores in diffusion models is equivalent to learning the noise (denoising)
- Recall the objective for learning the score network s

$$\mathcal{L}(\theta, \{\sigma_j\}_{j=1}^{L}) = \sum_{j=1}^{L} \sigma_j^2 \ell(\theta, \sigma_j), \qquad \ell(\theta, \sigma) = \mathbb{E}_{q(\tilde{x}|x)p_{\text{data}}(x)} \left[\frac{1}{2} \left\| s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q(\tilde{x}|x) \right\|^2 \right].$$

- Purification (deterministic update):

$$x_0 = x' + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$
$$x_t = x_{t-1} + \alpha_{t-1} s_\theta(x_{t-1}).$$

- 6 L Q F H W K H Q R U P R I L V E R X Q G H G G X H W R L Q G L V W L Q J X L V K D E L O L W \ F R Q V W U D L Q W W K U H O D W L Y H O \ V P D O O S H U W X U E D W L R Q
- The score network s is trained to denoise images perturbed by Gaussian noises. Adding Gaussian noises makes x_0 more similar to the data used to train the score network.

Figure 1. An illustration of *DiffPu*

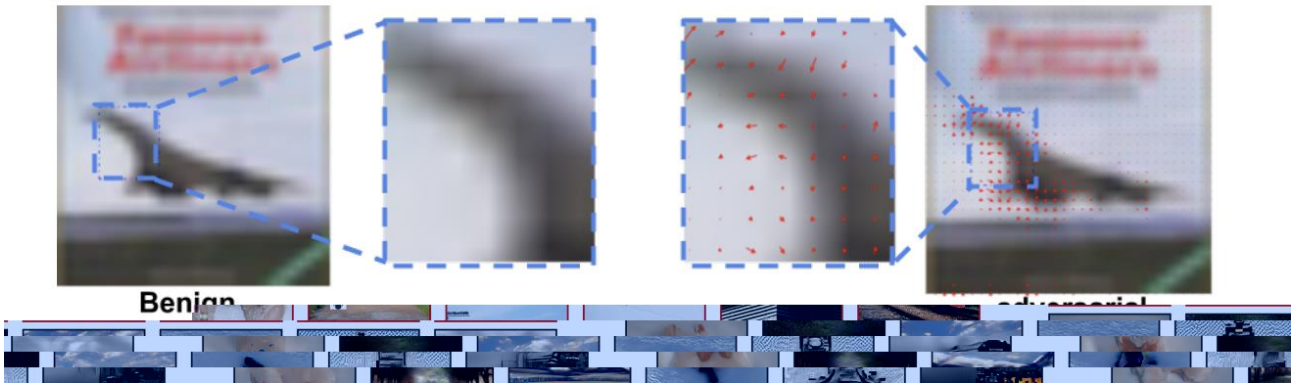
e. Given a pre-trained diffusion model, we add noise to adversarial images following the forward diffusion process with a small diffusion timestep t^* to get diffused images, fr
m which we recover clean images through the reverse denoising process before classific
tion. Adaptive att

Table 4. Standard accuracy and robust accuracies against unseen threat models on ResNet-50 for CIFAR-10. We keep the same evaluation settings with (Laidlaw et al., 2021), where the attack bounds are $\epsilon = 8/255$ for AutoAttack ℓ_∞, $\epsilon = 1$ for AutoAttack ℓ_2, and $\epsilon = 0.05$ for StAdv. The baseline results are reported from the respective papers. For our method, the diffusion timestep is $t^* = 0.125$.

Method	Standard Acc	Robust Acc

[redacted]

StAdv



Motivations

‡ A new method based on a diffusion model (DM) to reconstruct images from human brain activity obtained via functional magnetic resonance imaging (fMRI)

‡ Reconstructing visual images from fMRI is challenging:

- ± The underlying representations in the brain are largely unknown
- ± The sample size typically associated with brain data is relatively small
- ± Low signal-to-noise ratio with fMRI data

‡ Overarching goal:

- ± use DMs for high resolution visual reconstruction
- ± use brain encoding framework to better understand the underlying mechanisms of DMs and its correspondence to the brain.

fMRI

fMRI measures brain activity by detecting changes associated with blood flow.

When an area of the brain is in use, blood flow to that region also increases

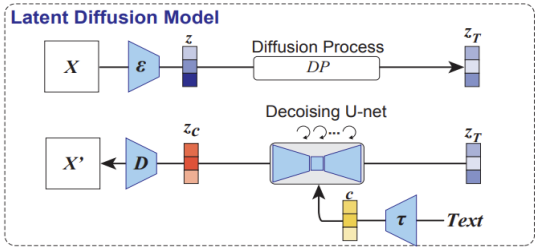
Noises in fMRI:

1. thermal noise

Data

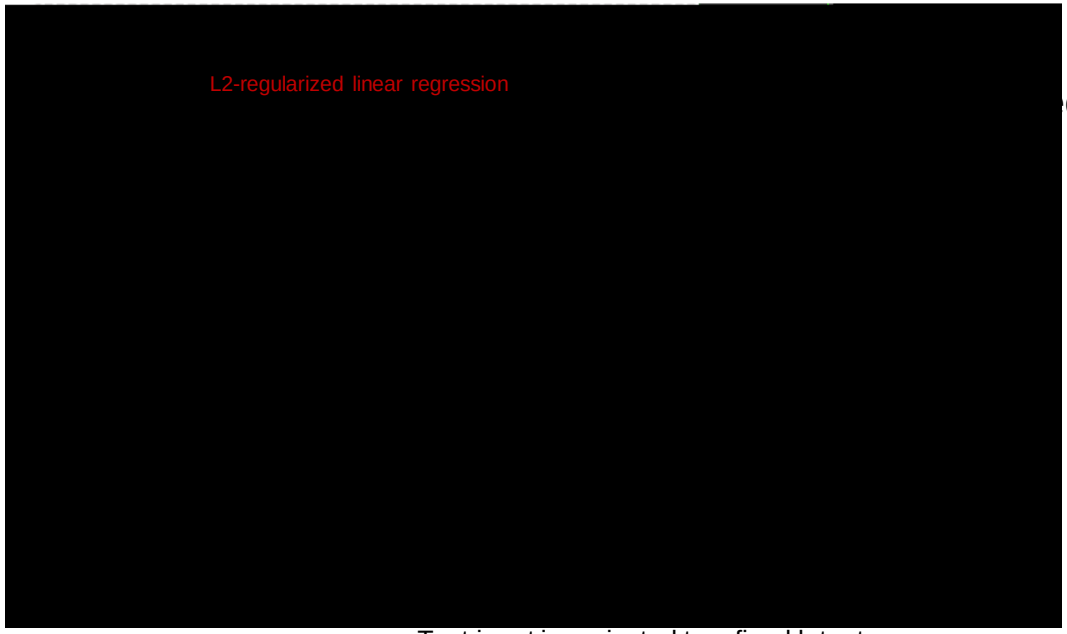
- Natural Scenes Dataset (NSD)
- NSD provides data acquired from a 7-Tesla fMRI scanner over 30 ±40 sessions during which each subject viewed three repetitions of 10,000 images.
- The images used in the NSD experiments were retrieved from MS COCO and cropped to 425*425

Framework



The only training required in our method is to construct linear models that map fMRI signals to each LDM component

z as the latent representation of the original image compressed by the autoencoder



coded image Xz

To construct models from fMRI to the components of LDM, we used L2-regularized linear regression, and all models were built on a per subject basis

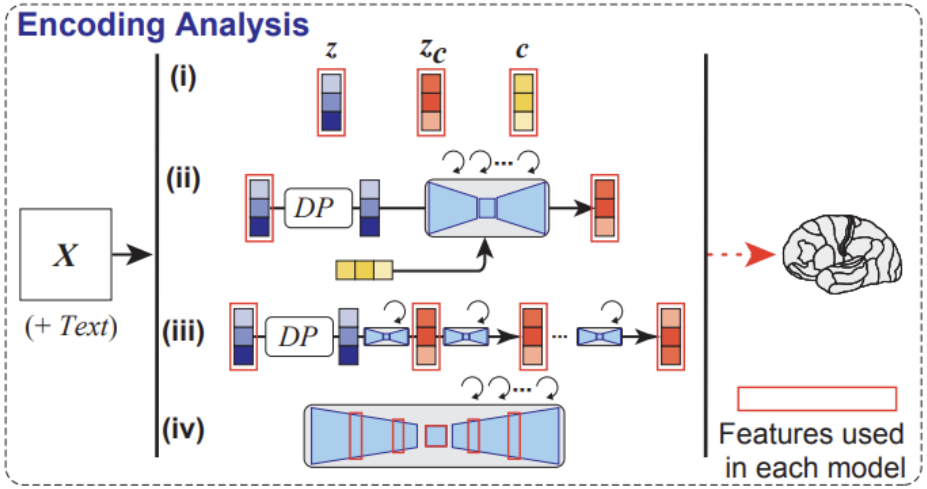
Text input is projected to a fixed latent representation by a pretrained text encoder (CLIP)

We decoded latent representations of the presented image (z) and associated text c from fMRI signals within early (blue) and higher (yellow) visual cortices, respectively.

These latent representations were used as input to produce a reconstructed image Xzc.

Encoding: Whole-brain Voxel-wise Modeling

Encoding: Whole-brain Voxel-wise Modeling
map representations in LDM to brain activity



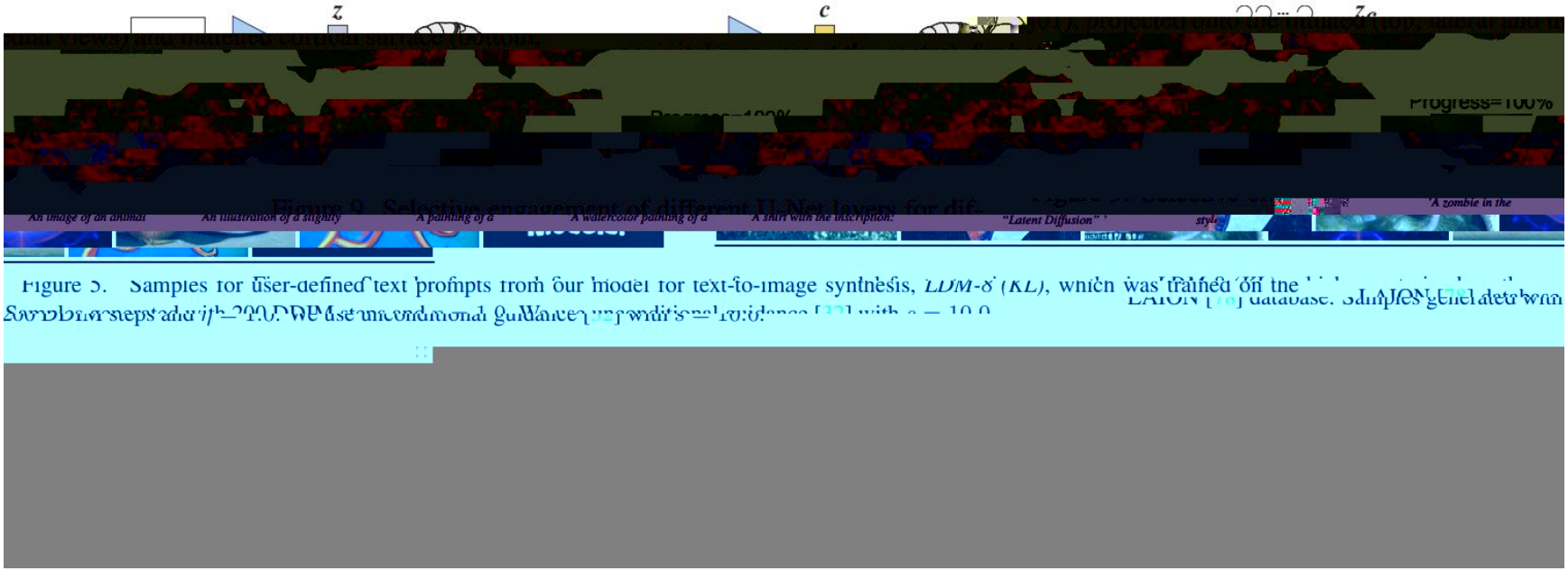
Schematic of encoding analysis. We built encoding models to predict fMRI signals from different components of LDM, including z, c, and zc.

(i) linear models to predict voxel activity from latent representations independently: z, c, and zc
(ii) we incorporated them into a single model
(iii) examine how zc changes through the denoising process.
(iv) we extracted features from different layers of U-Net

z produced high prediction performance in the posterior part of visual cortex, namely early visual cortex

c produced the highest prediction performance in higher visual cortex.

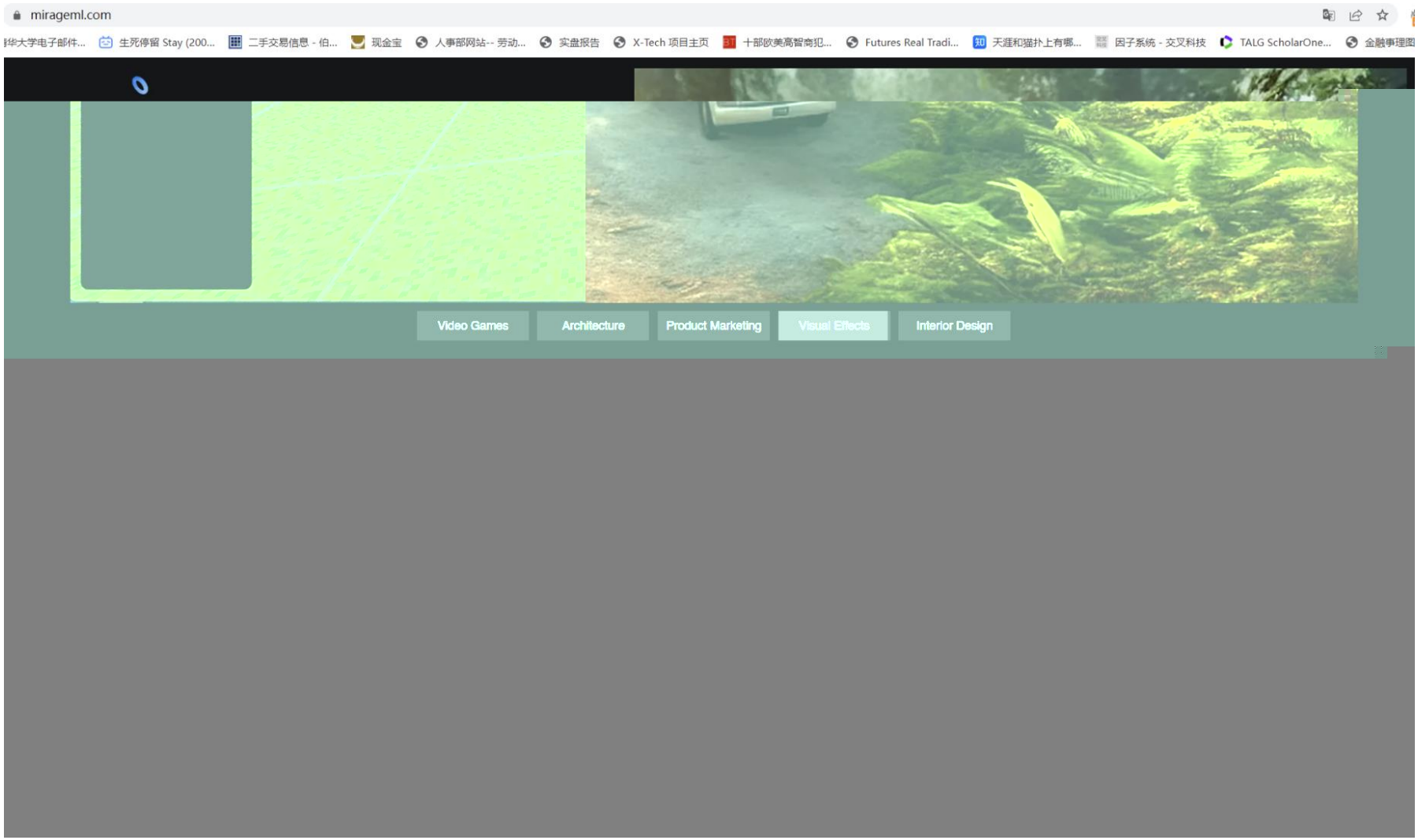
zc carries a representation that is very similar to z, showing high prediction performance for early visual cortex



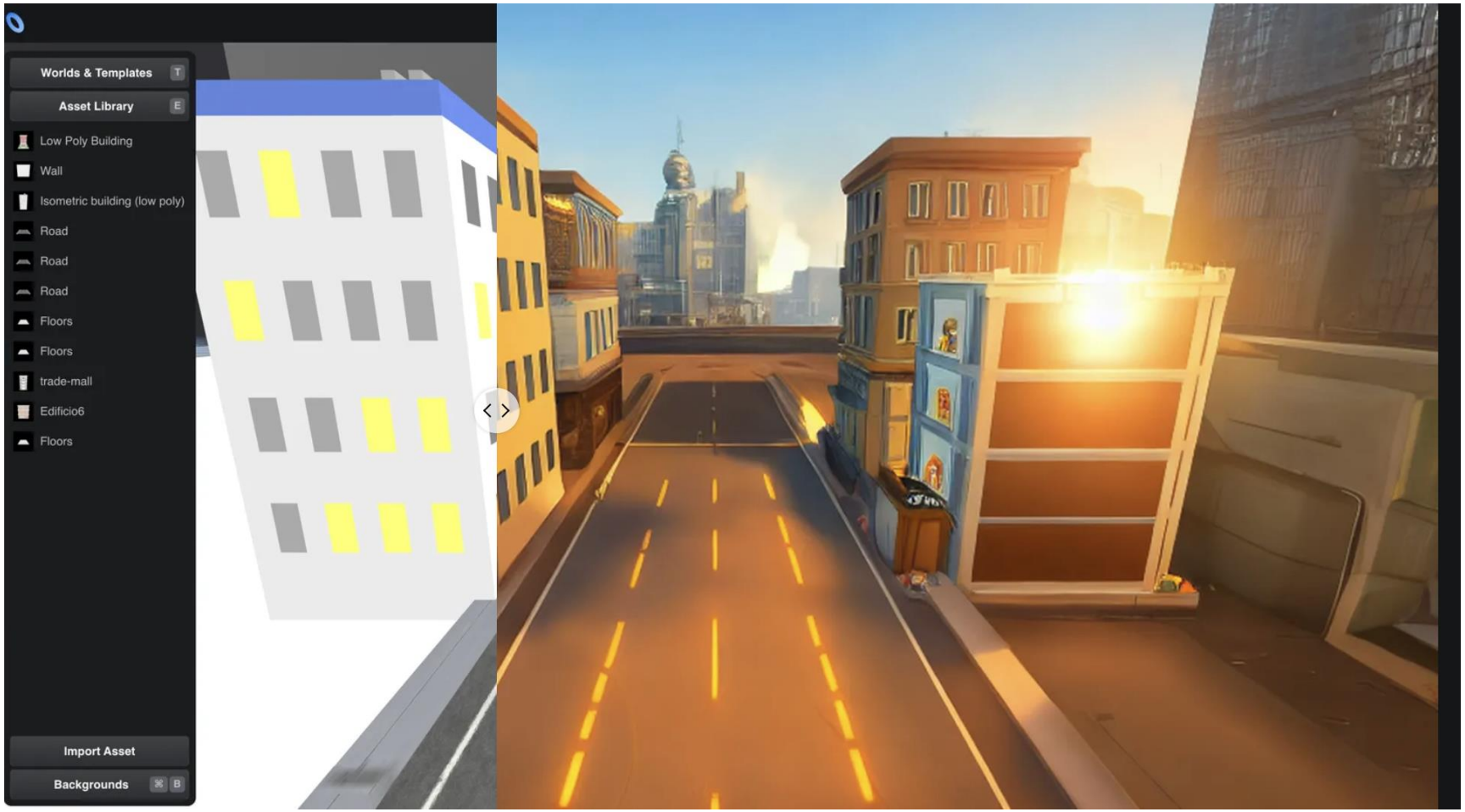
Stable Diffusion: Layout-to-Image

Figure 8. Layout-to-image synthesis with an *LDM* on COCO [4], see Sec. 4.3.1. Quantitative evaluation in the supplement D.3.

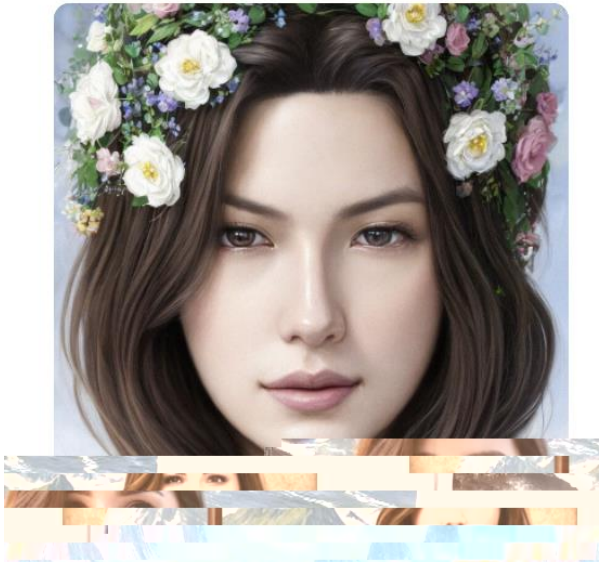
3D design


miragemI.com
Video Games
Architecture
Product Marketing
Visual Effects
Interior Design

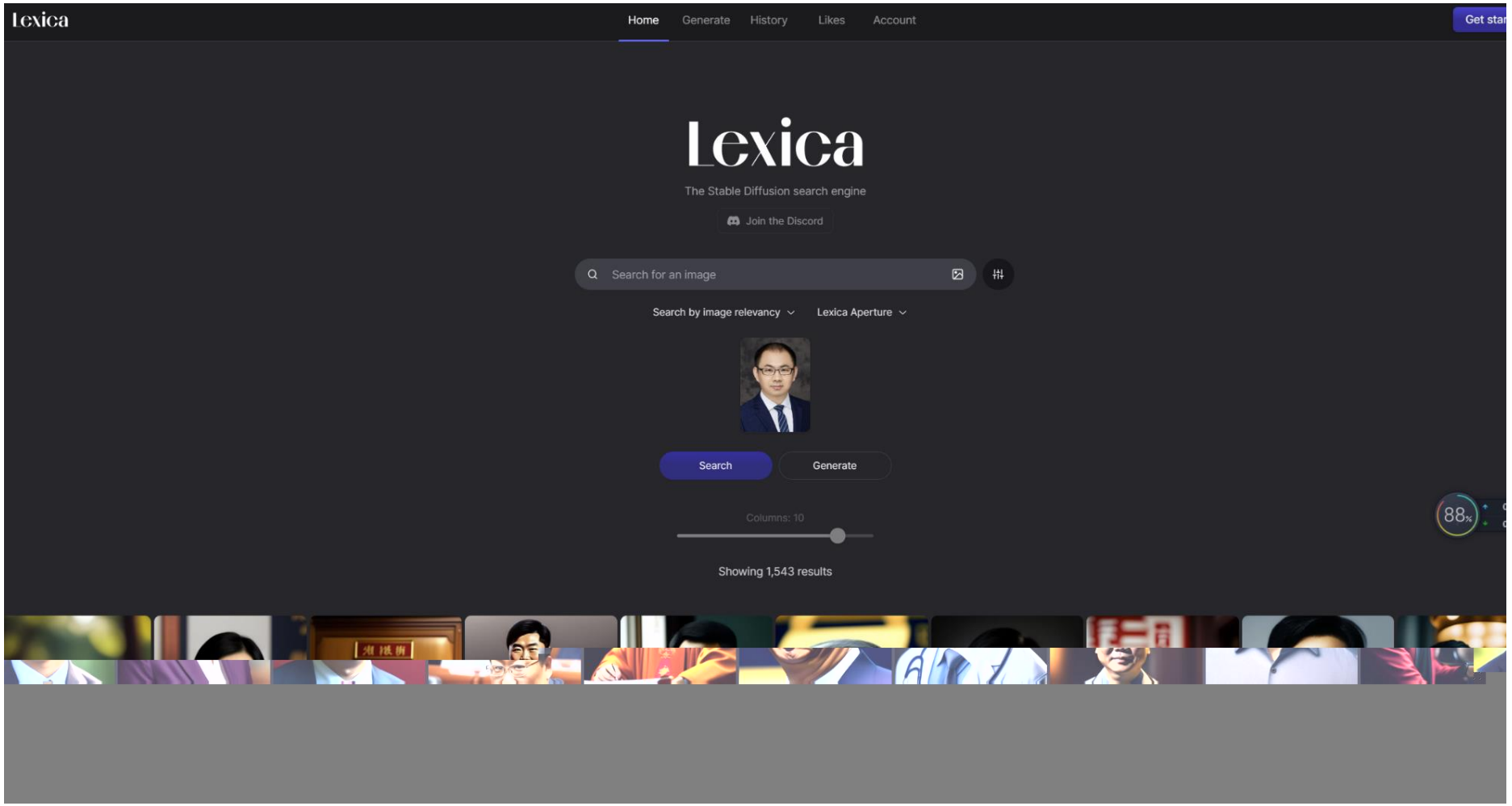
3D design


Worlds & Templates T
Asset Library E
Low Poly Building
Wall
Isometric building (low poly)
Road
Road
Road
Floors
Floors
trade-mall
Edificio6
Floors
Import Asset
Backgrounds ⌘ B

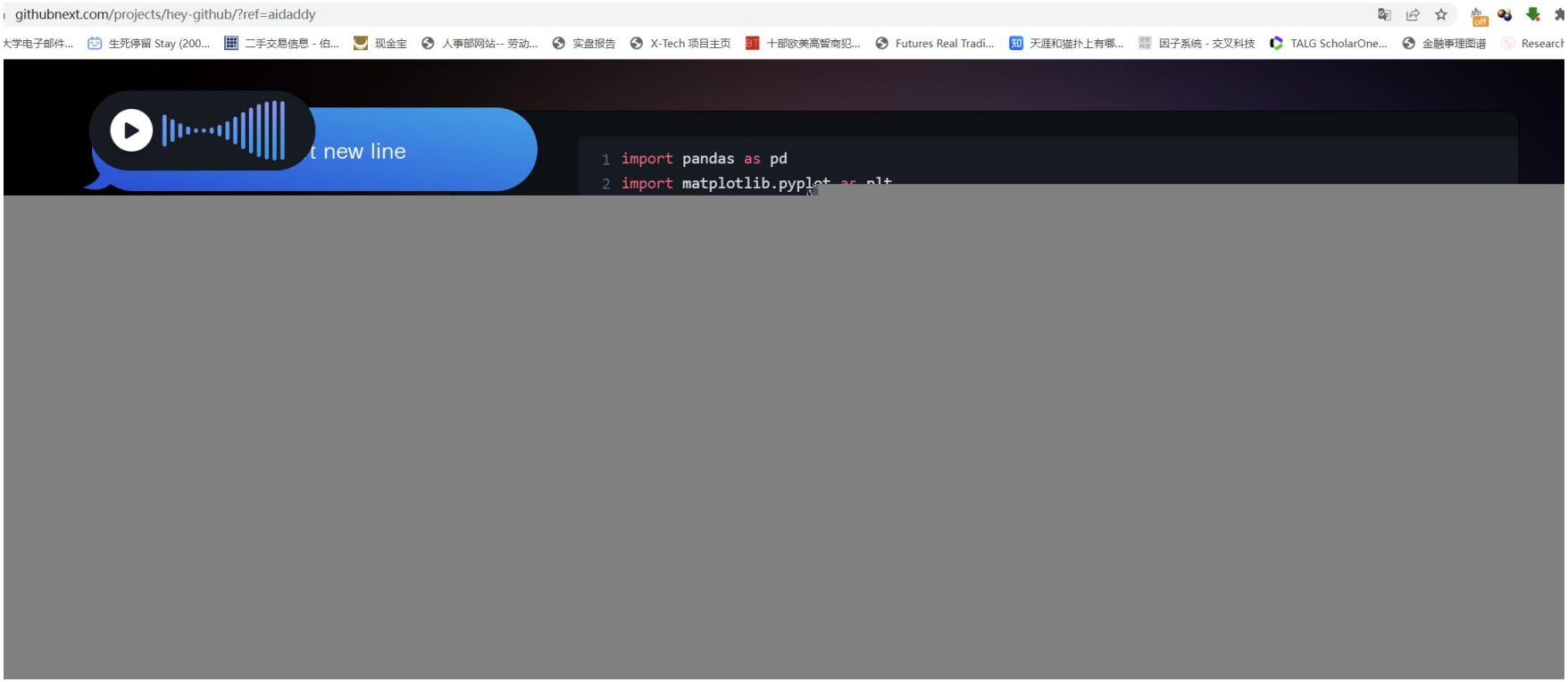
https://neural.love/



Stable diffusion search engine



AIGC: Low-code/No-code

‡Write python without touching the keyboard

‡ https://githubnext.com/projects/hey-github/?ref=aidaddy