Exploring Open Domain Image Super-Resolution through Text

Paramanand Chandramouli, Kanchana Vaishnavi Gandikota
University of Siegen

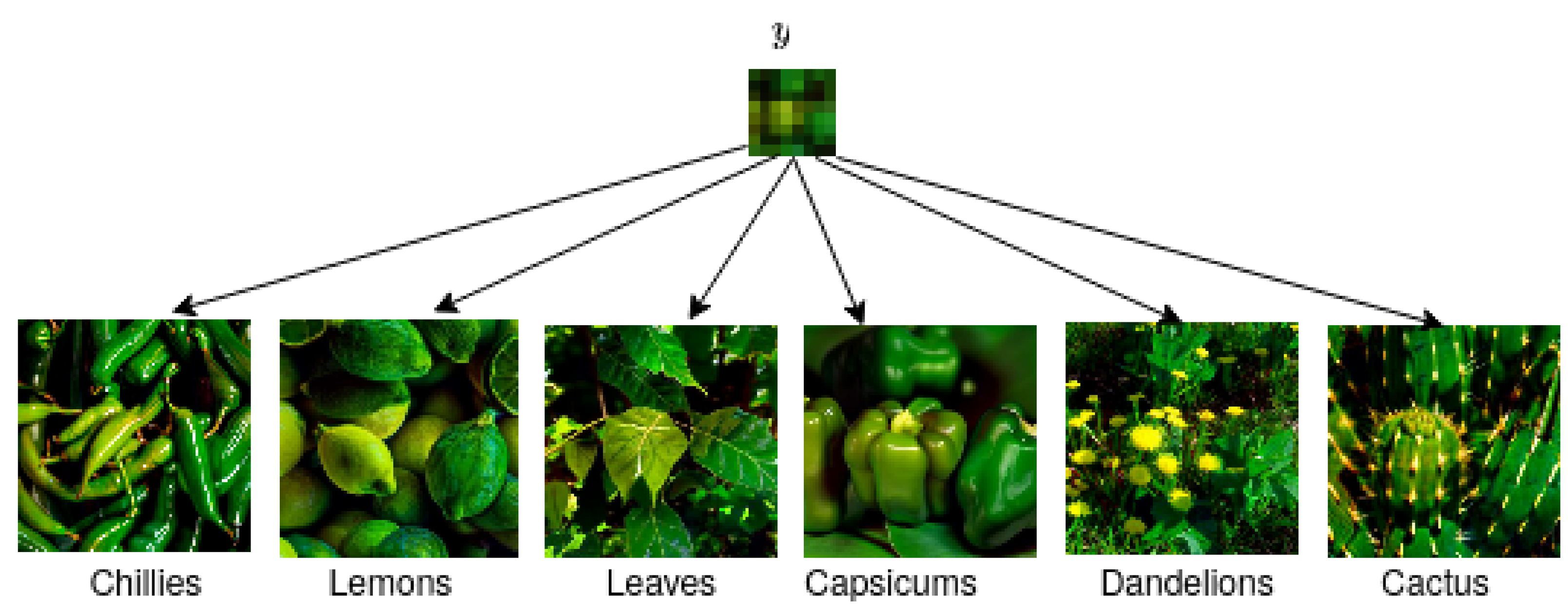
Introduction

• Image super-resolution (SR) aims to recover a high resolution (HR) image from a low-resolution (LR) input $\mathbf{y}$.

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$$

• SR is highly ill-posed – many valid solutions satisfy data consistency accurately.

Our Goal Explore multiple consistent solutions through text.



Our Solution Adapt text-to-image diffusion model ***DALLE2-unCLIP*** for SR by analytically enforcing consistency of the solutions with the input LR image for diverse text inputs.

Preliminaries

Range space-null space decomposition (RND)– useful to construct a consistent solution $\hat{\mathbf{x}}$ [3] from approximate solution $\bar{\mathbf{x}}$ to noiseless linear inverse problem $\mathbf{y} = \mathbf{A}\mathbf{x}$,

$$\hat{\mathbf{x}} = \underbrace{\mathbf{A}^{\dagger}\mathbf{y}}_{\text{range space}} + \underbrace{(\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A})\bar{\mathbf{x}}}_{\text{null space}}. \tag{1}$$

Wang et al.[2] modify the reverse diffusion process using RND. At time step t, estimate of clean image is given by

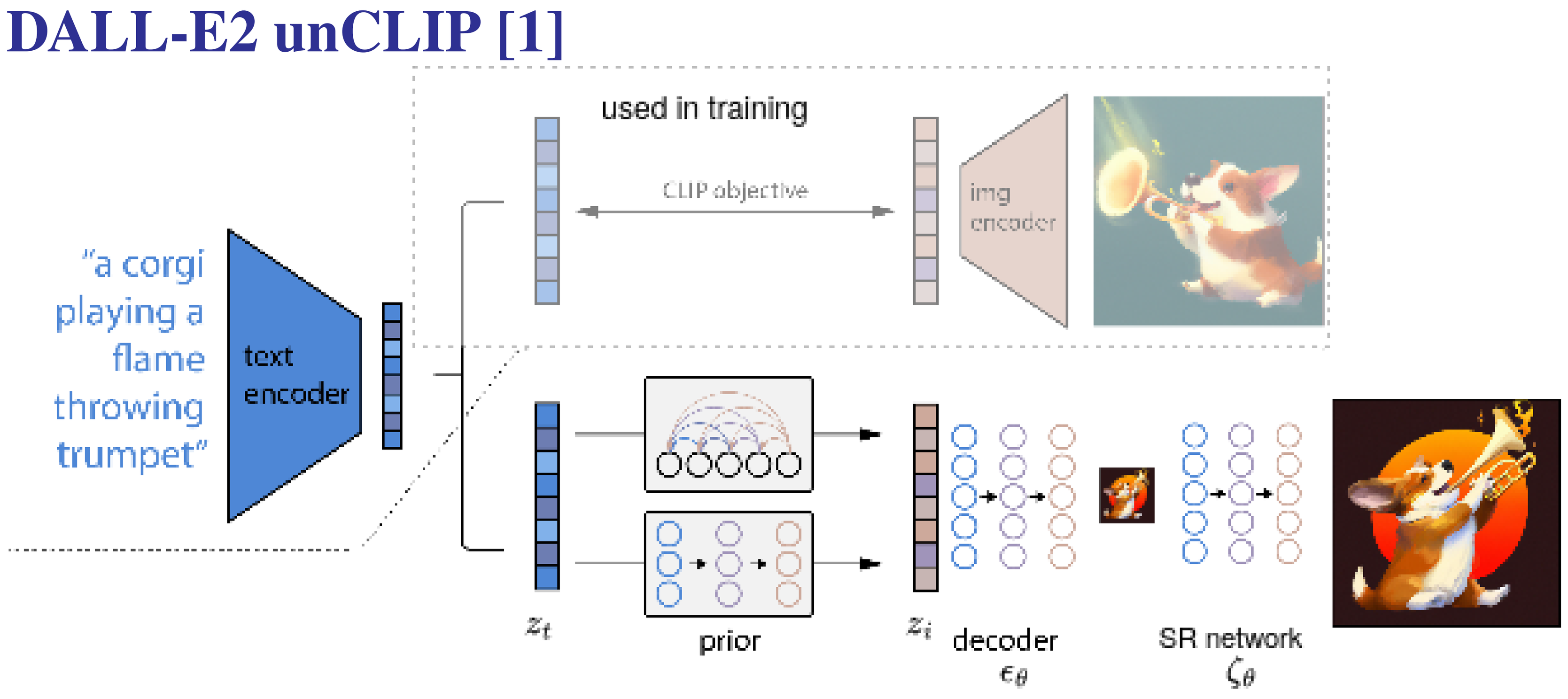
$$\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\sqrt{1-\bar{\alpha}_t}\right). \tag{2}$$

A rectified data consistent estimate $\hat{\mathbf{x}}_{0|t}$ is obtained from $\mathbf{x}_{0|t}$:

$$\hat{\mathbf{x}}_{0|t} = \mathbf{A}^{\dagger}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A})\mathbf{x}_{0|t}. \tag{3}$$

$\hat{\mathbf{x}}_{0|t}$ is used in subsequent sampling steps in reverse diffusion.

DALL-E2 unCLIP [1]



- A diffusion based prior to produce CLIP image embeddings z_i.
- A diffusion based generator $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ conditioned $\mathbf{z} = \{z_i, z_t\}$.
- A diffusion based SR module $\boldsymbol{\zeta}_{\boldsymbol{\theta}}$ to obtain a HR output.

Our Approach

Two Stage Consistency Enforcement

• Null space consistency in the reverse process of unCLIP decoder conditioned on $\mathbf{z} = \{z_i, z_t\}$ at each sampling step t:

$$\mathbf{x}_{LR_{0|t}} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_{LR_t} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_{LR_t}, t|\mathbf{z})\sqrt{1-\bar{\alpha}_t}\right)$$

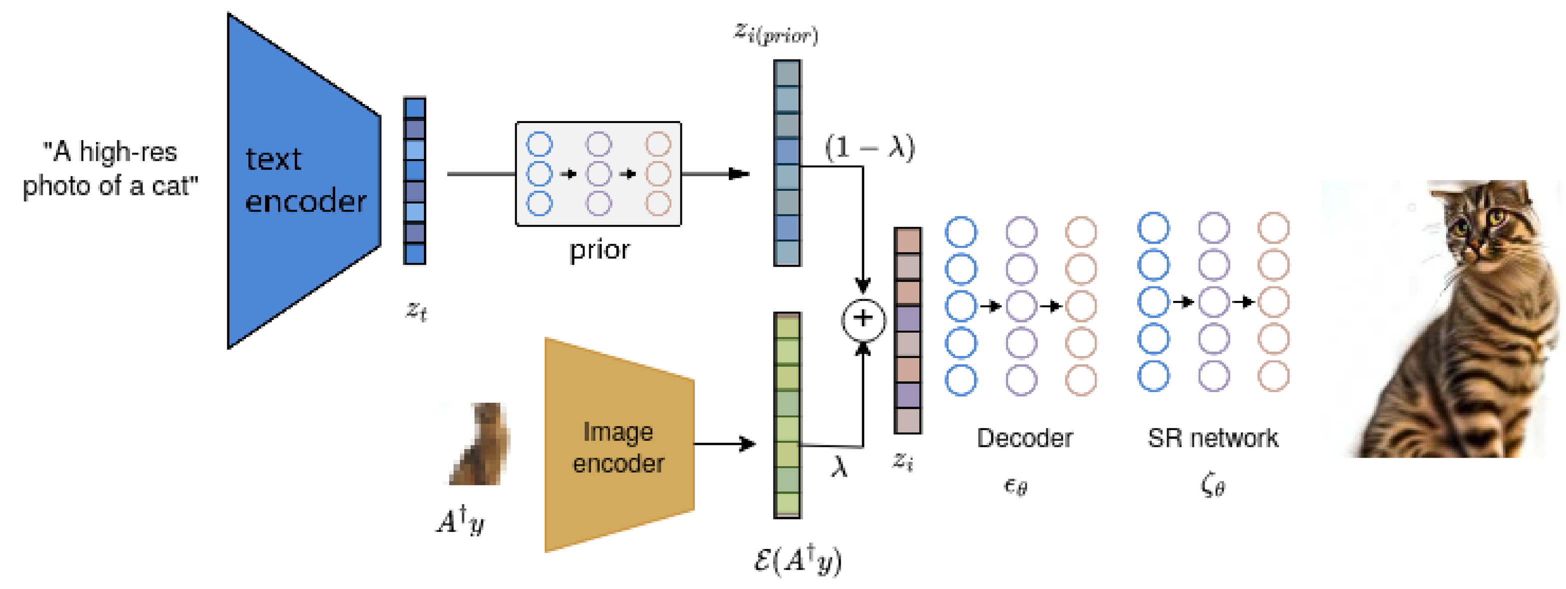
$$\hat{\mathbf{x}}_{LR_{0|t}} = \mathbf{A}^{\dagger}{}_{LR}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{\dagger}{}_{LR}\mathbf{A}_{LR})\mathbf{x}_{LR_{0|t}}$$

• Null space consistency in the reverse process of diffusion based SR module conditioned on $\mathbf{x}_{LR}$ at each sampling step t:

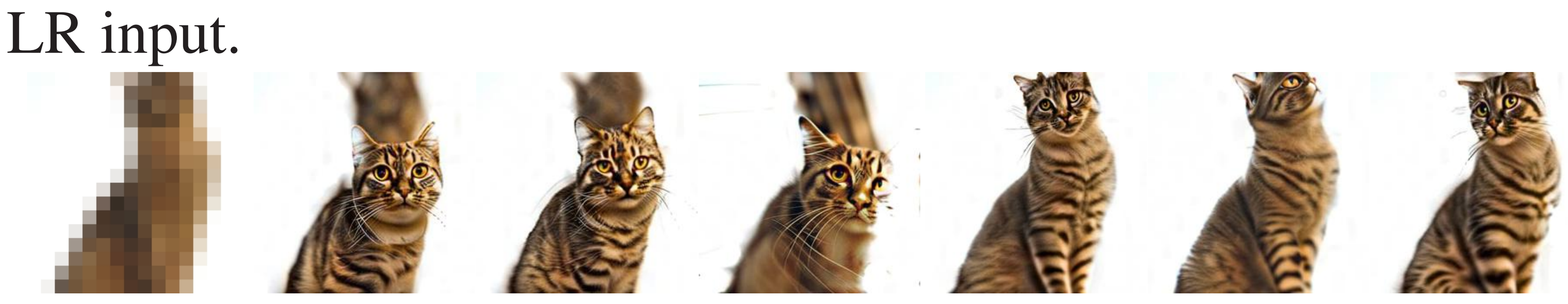
$$\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \boldsymbol{\zeta}_{\boldsymbol{\theta}}(\mathbf{x}_t, t|\mathbf{x}_{LR})\sqrt{1-\bar{\alpha}_t}\right)$$

$$\hat{\mathbf{x}}_{0|t} = \mathbf{A}^{\dagger}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A})\mathbf{x}_{0|t}$$

Embeddings Averaging Trick



• Improves structural consistency of the image embedding with the LR input.



Input LR ⟵ No averaging trick ⟶⟵Embeddings averaging ($\lambda = 0.4$)⟶

Results

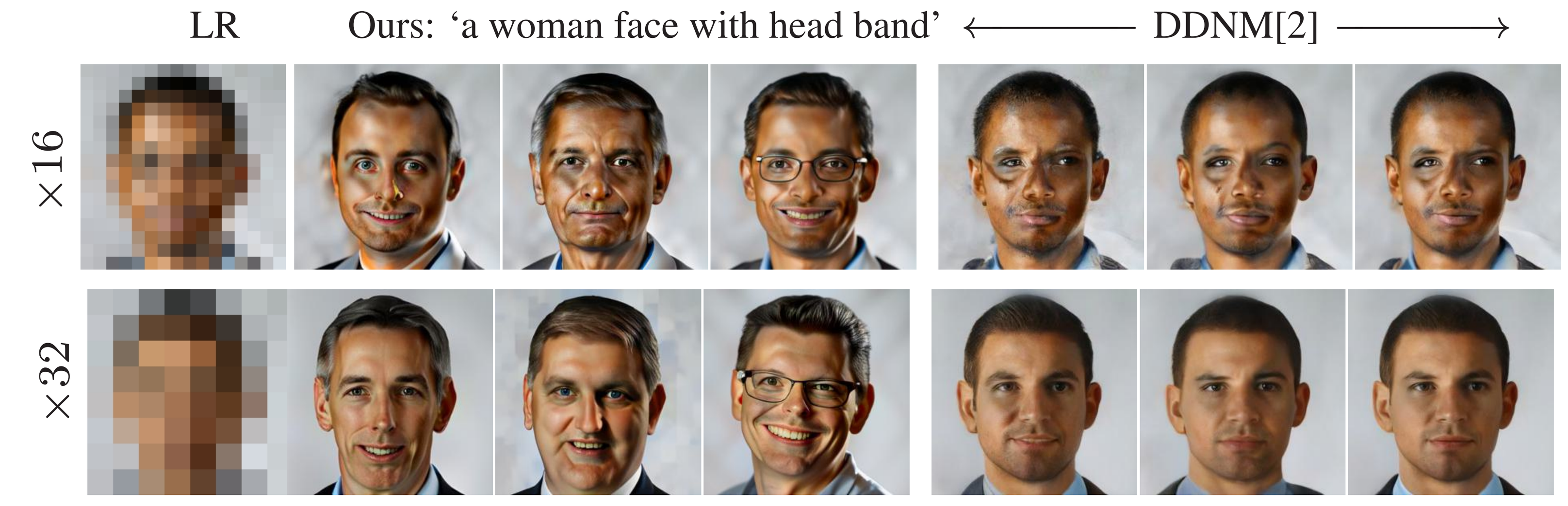
• Multiple consistent solutions for the same text prompt.

'A crown with blue jewels and diamonds' (16× SR)

'A fat blue jay sitting on a broken limb of a leafless tree' (16×SR)

LR Ours: 'a baby face with knitted cap' ⟵ DDNM[2] ⟶

LR Ours: 'a woman face with head band' ⟵ DDNM[2] ⟶



LR Ours: 'a man, man+grey hair, man+smile+glasses' ⟵ DDNM[2] ⟶

References

[1] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. In *NeurIPS*, 2022.

[2] Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023.

[3] Bahat, Y. and Michaeli, T. Explorable super resolution. In *CVPR*, 2020.