- 2024 Winter EECS542 -

Integrating Multimodal Techniques with Latent Diffusion Models to Advance Multi-View Optical Illusion Generation

Wonseok Oh
Umich ECE master's student
okong@umich.edu

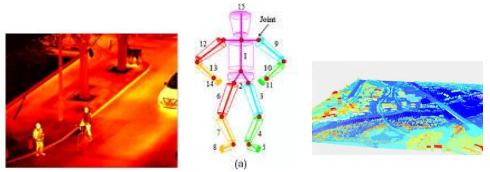
Yeheng Zong
Umich ECE master's student
yehengz@umich.edu

Contents

- Introduction
- Related works
- Method (raw model)
- Method (denoising)
- Results

Introduction

Hello, this is text.
Hahaha
I'm not robot. Apple computer is expensive

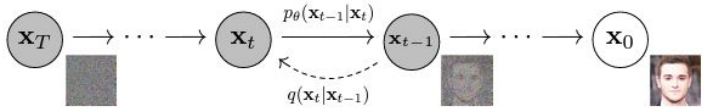


More modalities…

Limitations

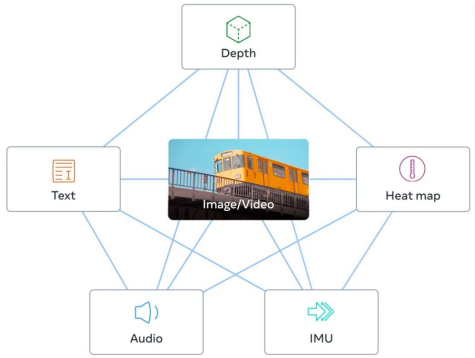
colab

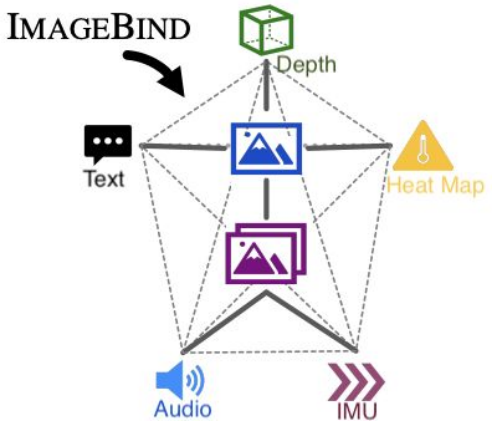
Related Works

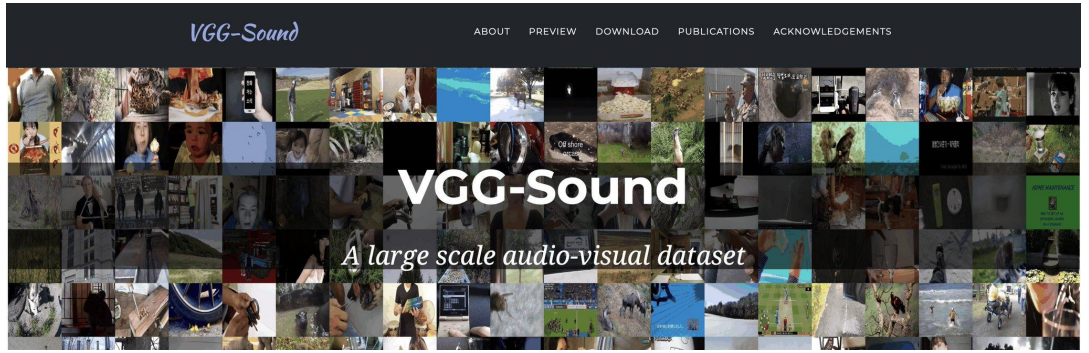


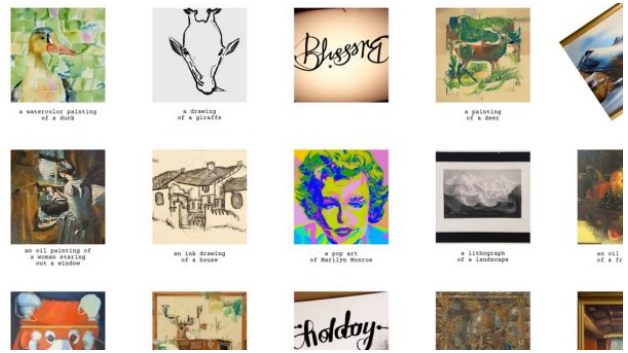
$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

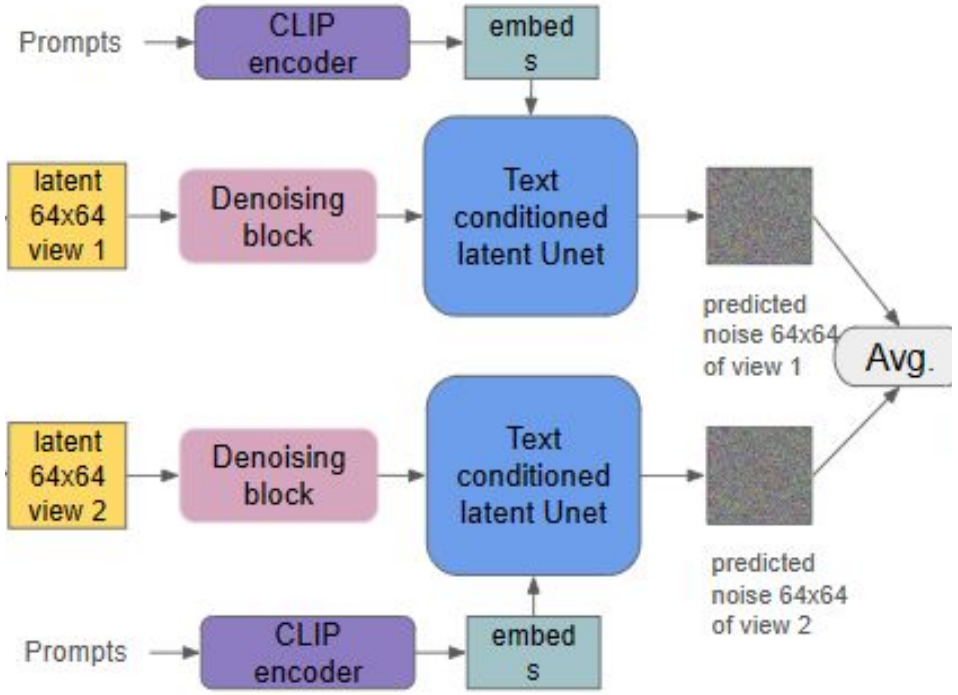




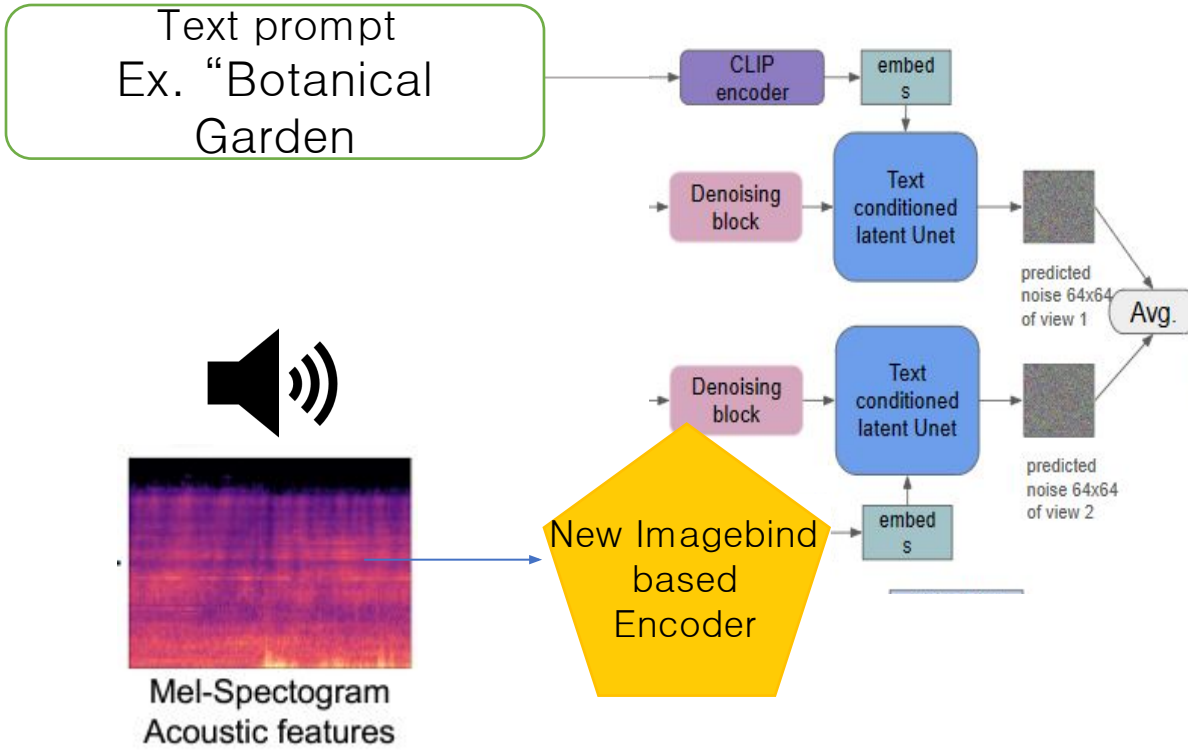




Method (Multimodal)



Raw model with two **Text**



Raw model with **Text** with **Sound**

Generation with text

an oil painting of waterfalls

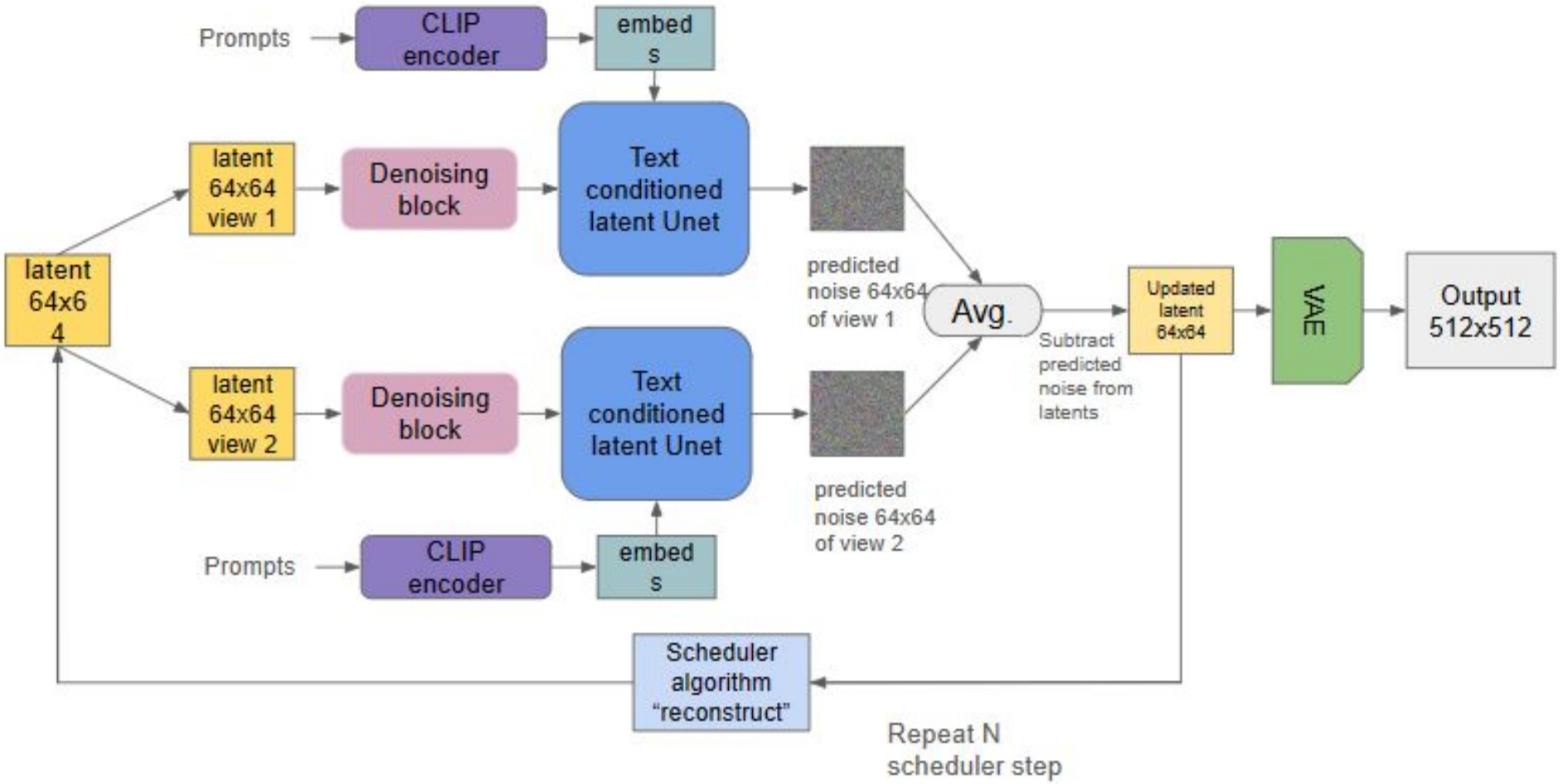
an oil painting of a dining table

Generation with sound

an oil painting of waterfalls(sound)

an oil painting of a dining table

Method (Overview)



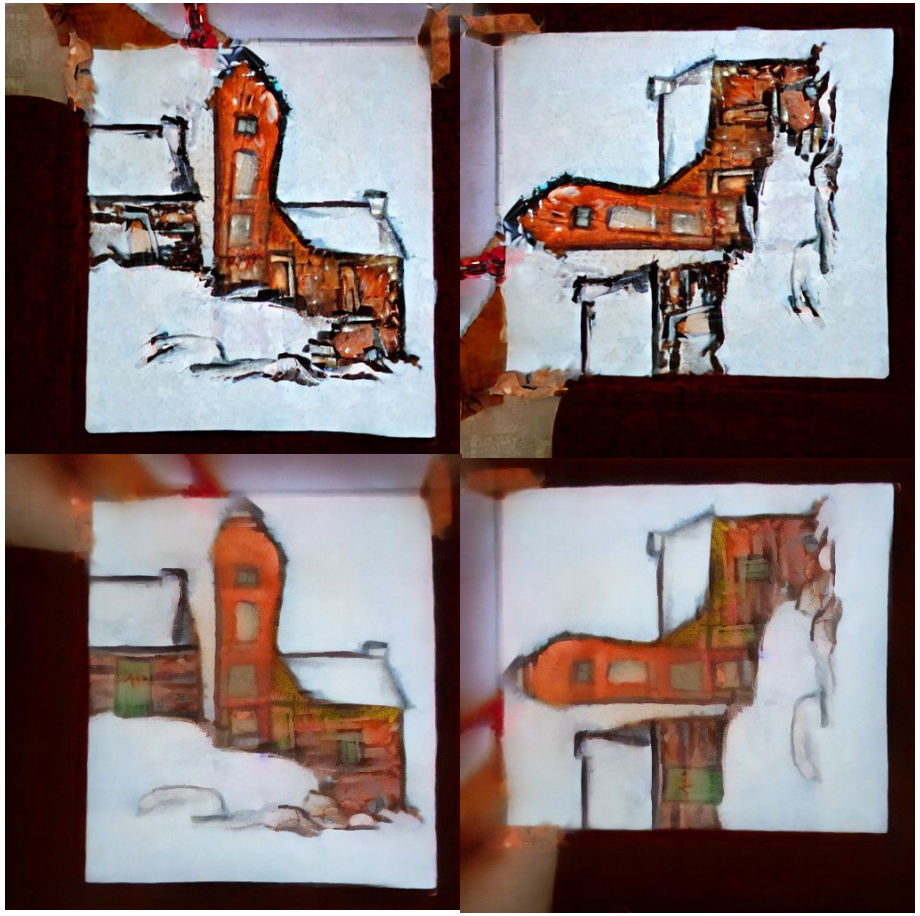
Method (Denoising)

- Fourier denoising
- Wavelet denoising
- Total Variance(TV) regularization

Fourier Denoising - Design

- Cut off percentage of fourier coefficient based on their magnitude
- Start Fourier denoising at step 200 and perform it every 20 steps
- Gradually decrease the scale of denoising

Fourier Denoising - Results



CLIP score	View 1	View 2
Raw generation	0.603	**0.754**
Fourier denoising generation	**0.736**	0.600

Prompt 1: A cartoon drawing of a snowy mountain village

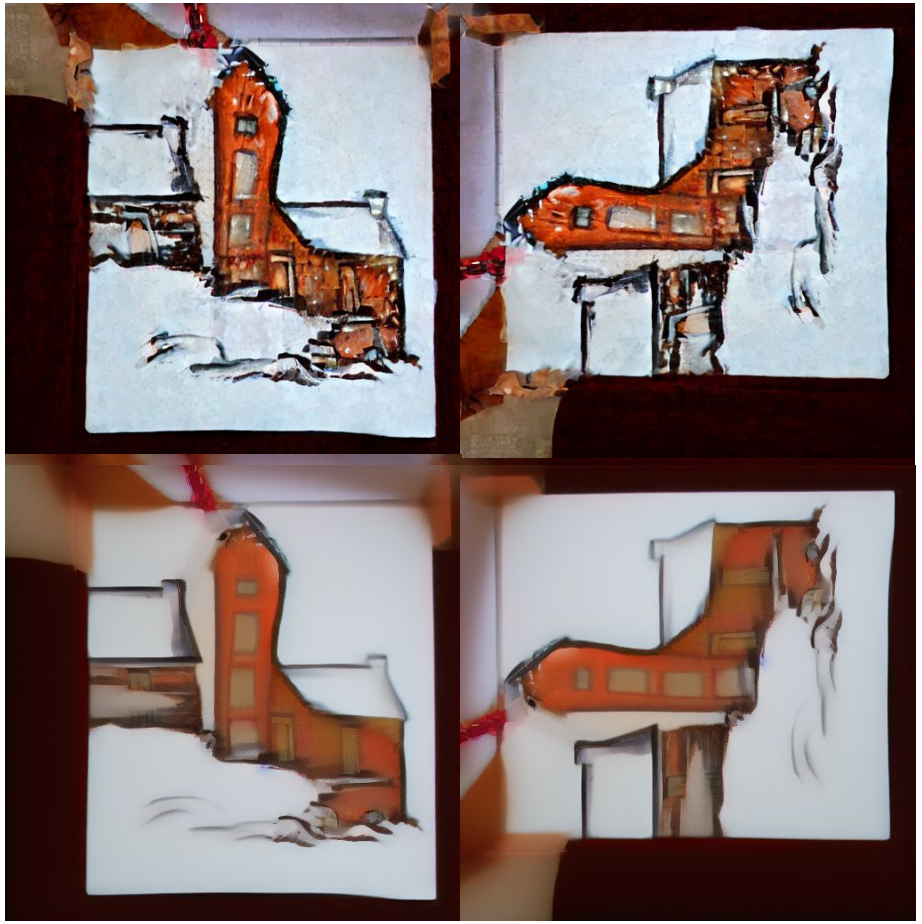
Prompt 2: A cartoon drawing of a horse

More Results

Wavelet Denoising - Design

- Apply soft-thresholding on the wavelet coefficients
- Start Wavelet denoising at step 200 and perform it every 20 steps
- Gradually increase the scale of denoising

Wavelet Denoising - Results



CLIP score	View 1	View 2
Raw generation	0.603	**0.754**
Fourier denoising generation	**0.716**	0.654

Prompt 1: A cartoon drawing of a snowy mountain village

Prompt 2: A cartoon drawing of a horse

TV regularization Denoising - Design

- Add TV as a regularizer
- Start TV regularization denoising at step 200 and perform it every 20 steps
- Gradually decrease the strength of regularization

TV regularization Denoising – Method

The objective function for least squares with TV regularization can be expressed as:

$$\min_x \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda \cdot \mathrm{TV}(x) \tag{1}$$

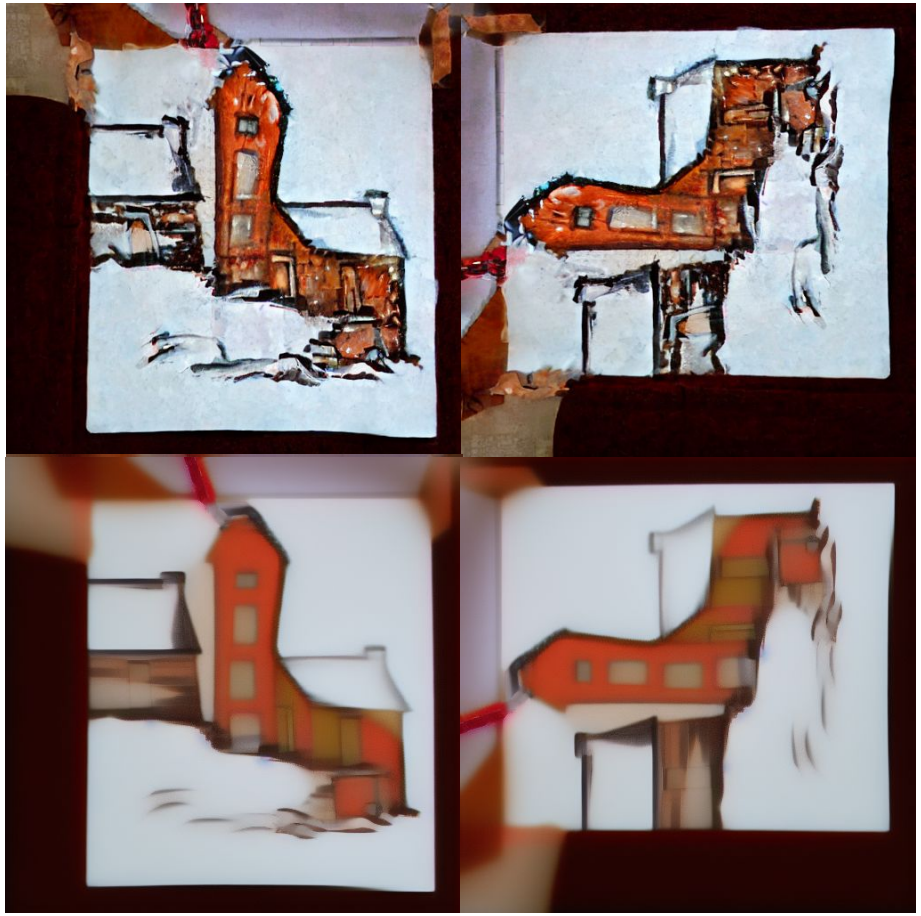
$$\text{where} \quad \mathrm{TV}(x) = \sum_{i,j} \sqrt{|x_{i+1,j} - x_{i,j}|^2 + |x_{i,j+1} - x_{i,j}|^2} \tag{2}$$

$$\text{in practice, we use} \quad \mathrm{TV}(x) = \sum_{i,j} (\sqrt{|x_{i+1,j} - x_{i,j}|^2} + \sqrt{|x_{i,j+1} - x_{i,j}|^2}) \tag{3}$$

Here,

- A is the system matrix.
- x is the image (or signal) to be reconstructed.
- b is the observed data.
- λ is the regularization parameter that controls the trade-off between the fidelity to the data and the smoothness of the solution.
- $\mathrm{TV}(x)$ is the total variation of x, promoting sparsity in the gradient of the image, thus preserving edges while smoothing.

TV regularization Denoising - Results

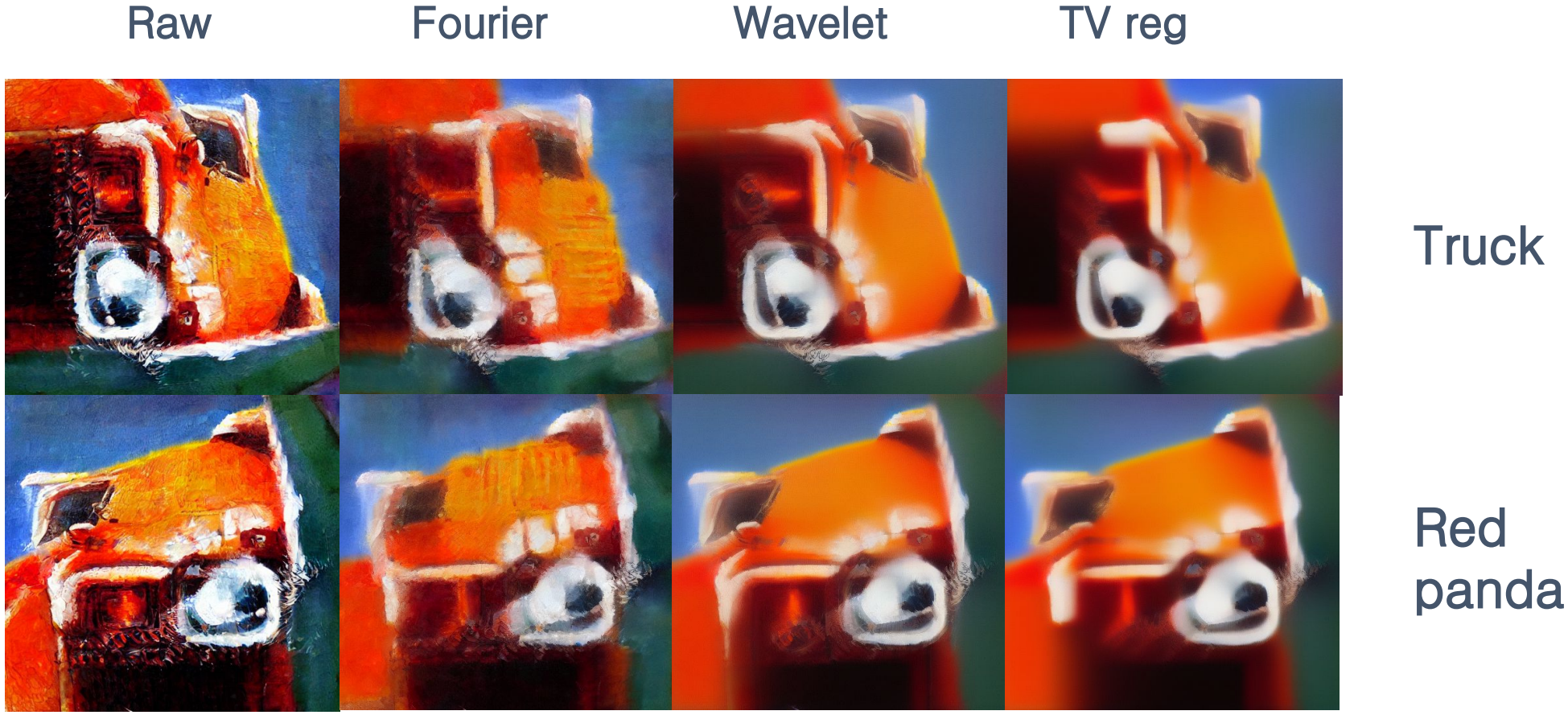


CLIP score	View 1	View 2
Raw generation	0.603	0.754
Fourier denoising generation	**0.756**	**0.833**

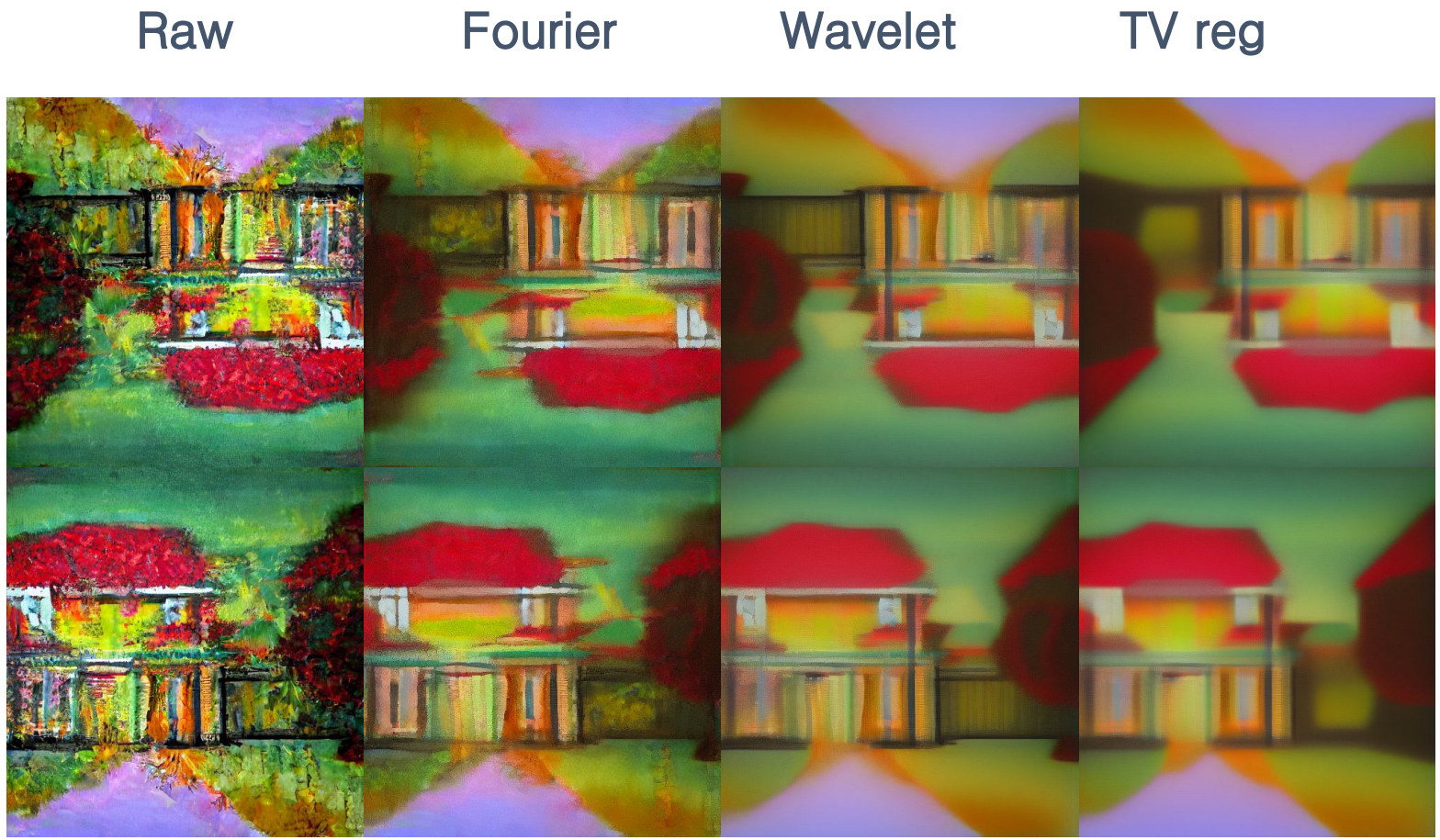
Prompt 1: A cartoon drawing of a snowy mountain village

Prompt 2: A cartoon drawing of a horse

Discussion of denoising block



Discussion of denoising block



Future works

- Make more modalities
- Generate more general, apply to various cases
- Use other image denoising technique

Thank you

Reference

[1] Geng, D., Park, I., & Owens, A. (2023). Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models. *arXiv preprint arXiv:2311.17919*.

[2] Boigné, E., Parkinson, D. Y., & Ihme, M. (2022). Towards data-informed motion artifact reduction in quantitative CT using piecewise linear interpolation. *IEEE Transactions on Computational Imaging*, *8*, 917-932.

[3] Girdhar, Rohit, et al. "Imagebind: One embedding space to bind them all." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[4] Chen, Honglie, et al. "Vggsound: A large-scale audio-visual dataset." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.