Peyresq 2026 Summer School - PhD presentations

Multimodal Diffusion Model for High-Resolution Earth Observation Data

Victor Sanchez

22-06-2026

CESBIO, Université de Toulouse, ISAE-SUPAERO, Toulouse, FRANCE

Table of Content

Context

Context - Essential Climate Variables (EVs)

The Global Climate Observing System defined Essential Climate Variables (EVs) about the atmosphere, cryosphere, ocean, anthroposphere, biosphere, and the hydrosphere.
Monitoring EVs provides key information to **understand climate, biodiversity and environmental changes** using Earth Observation (EO) systems.

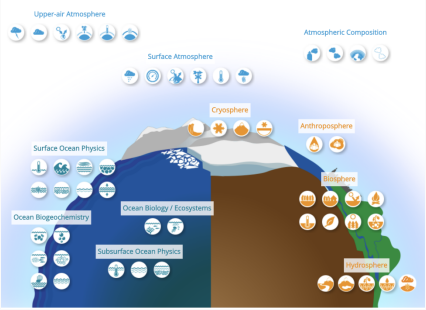


Figure 1:[1]

[1] https://www.earthdata.nasa.gov/learn/earth-observation-data-basics/essential-variables.

Context - Earth Observation

Retrieving EVs from multi-source data is challenging due to :

- The singularities of EO data (ex: indirect observation of interest variables);
- Varying spatial resolution;
- Irregularly sampled time series

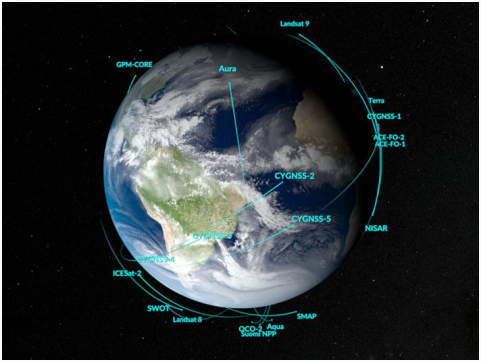


Figure 2: Variability of satellites orbits.[a]

[a] https://www.earthdata.nasa.gov/.

Context - Deep Learning

Figure 3: AI generated image from[a]

- Learn complex patterns from huge amounts of data.

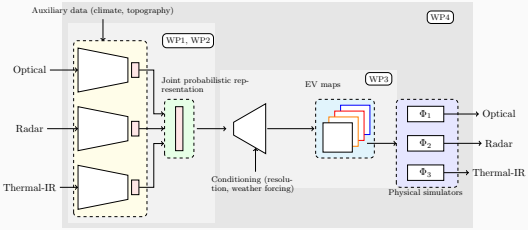
- Recent models lack multi-modality diversity. A majority only takes only Sentinel 2 and/or Sentinel 1 when there are much more data available (LANDSAT, MODIS, etc ...).

- Training also requires large amounts of labeled data, which are scarce and noisy in EO. Consequently, current models have a restricted usage in large scale EO systems.

[a] https://pixabay.com.

RELEO Project - ANITI Chair

RELEO Project - Existing Work and Objectives of Project



- New self-supervised representation learning [Dumeur, Valero, and Inglada 2024b] methods.

- Produce semantically meaningful probabilistic representations [Dumeur, Valero, and Inglada 2024a] from high-dimensional multi-modal EO data.

- Usage of prior knowledge from physical models into Deep Learning (DL).

- Process multi-modal data to capture complementary spatio-temporal patterns.

- Condition the decoding of generic embeddings to retrieve and forecast EVs and their uncertainties.

Why choosing Perceiver-IO [Jaegle et al. 2021] ?

- No need for different encoding heads for different kinds of data
- Leverages cross-attention to reduce the cost of the deeper layers of the network
- Handle general purpose inputs and outputs while scaling linearly in both input and output size

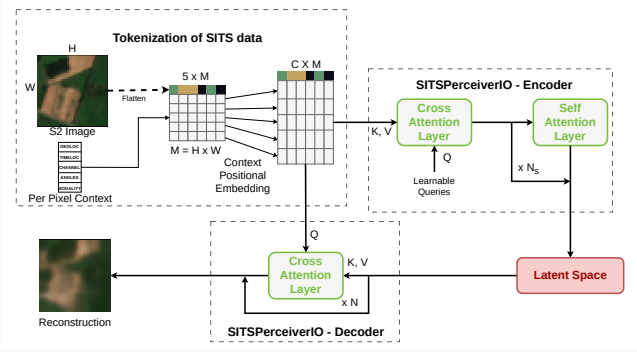


Figure 4: Perceiver-IO [Jaegle et al. 2021] example for Auto-encoding task

RELEO Project - Opportunities of Probabilistic Generative Approach

Flexible probabilistic generation depending on available data.

Probabilistic Multi-modal data fusion.

Conditioning on auxiliary data for downstream tasks.

Estimation of EVs with uncertainty.

Probabilistic Generative Approach

Diffusion Model - Forward / Backward Process

Forward Process : Gradually converts the observed data x_0 into a noisy version x_T by passing the data throught T steps of a stochastic encoder $q(x_t|x_{t-1})$ until we have $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

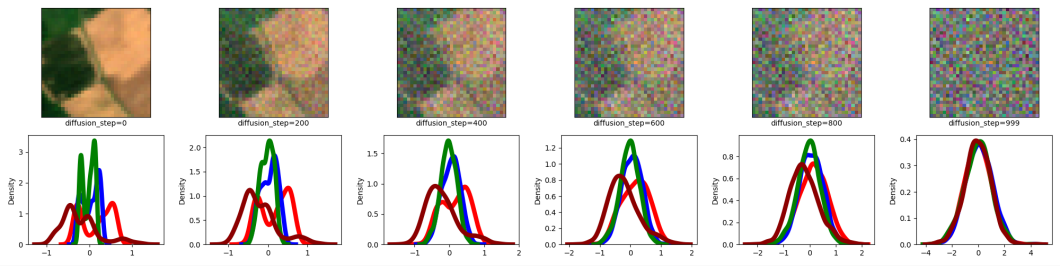


Figure 5: Forward Process with Cosine Schedule. Effect on raw data (middle) and normalized data (bottom

Backward Process : Learn parameters θ to perform a reverse process to undo the forward process by passing the noisy image through T steps of a decoder $p_\theta(x_{t-1}|x_t)$ until we recover x_0.

Diffusion Models - Motivations

"***The basic idea behind these models is based on the observation that it is hard to convert noise into structured data, but it is easy to convert structured data into noise.***" Kevin P. Murphy in [Murphy 2023]

Diffusion Models is a group of Probabilistic Approach known for :

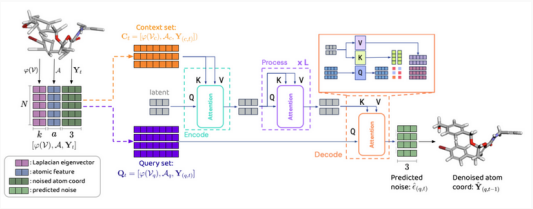
- Its ability to **generate diverse, high quality, samples** [Rombach et al. 2022]

- Its relative **simplicity of the training scheme**, which allows very large models to be trained at scale [Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2022]

- Its relative long process of generation steps, but already tackled by several improvements (DDIM, DPM, DPM++, EDM) [Song, Meng, and Ermon 2022,Karras, Aittala, Lehtinen, et al. 2024, Karras, Aittala, Aila, et al. n.d., Lu, Zhou, J. Chen, et al. n.d., Lu, Zhou, Bao, et al. 2025]

Doing Diffusion Model With Perceiver IO?

Apparently it is possible to perform diffusion with perceiver IO as a backbone! Some recent papers use a Perceiver-IO for doing diffusion !



Application on Molecule structural generation : [Wang et al. 2024]

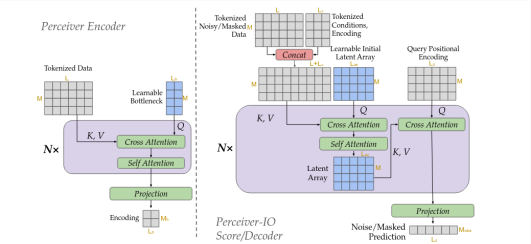


Figure 2: Schematic of the perceiver encoder and perceiver-IO score/decoder model used in daep and mae.

Application on Astronomy : [Shen and Gagliano 2025]

Application to Earth Observation Data

Diffusion Models applied to EO - State of the Art

State of the Art using Diffusion Models for Earth Observation:

Name	DiffusionSat[2]	CRS-Diff[3]	COP-GEN[4]
Latent or Pixel Space	Latent with pretrained SD	Latent with pretrained SD	Latent with pretrained SD
Modalities	Sentinel-2 (RGB)	text, Sentinel-2 (RGB)	Sentinel-2, Sentinel-1, DEM, LULC
Temporal Correlation	Multi-Temporal on demand	None	None

[2] Samar Khanna et al. (2023). **"Diffusionsat: A generative foundation model for satellite imagery."** In: *arXiv preprint arXiv:2312.03606*.

[3] Datao Tang et al. (Sept. 2024). ***CRS-Diff: Controllable Remote Sensing Image Generation with Diffusion Model.*** en. arXiv:2403.11614 [cs]. DOI: 10.48550/arXiv.2403.11614. URL: http://arxiv.org/abs/2403.11614 (visited on 10/06/2025).

[4] Miguel Espinosa et al. (2026). **"COP-GEN: Latent Diffusion Transformer for Copernicus Earth Observation Data–Generation Stochastic by Design."** In: *arXiv preprint arXiv:2603.03239*.

Multimodality and Diffusion model ?

The majority of generative models tends to learn and generate $p(x|y)$ but using a *joint model* means being able to learn and generate $p(x, y)$. x and y being different modalities.

***Multimodality* in diffusion based architecture :**

- Usually from text to image (**which is a conditioned model and not joint model**)
- Essentially using RGB images
- Neither multispectral nor multi-source inputs for remote sensing.

Be careful, in computer vision, **Multimodal** does not imply **Joint**!
But **Joint** does not imply **Multimodal**!
In our case, we want a joint and multimodal diffusion model !

A work from [Bao et al. 2023a], showed that it is possible to handle multiple modalities with a single diffusion model to perform : Unconditional sampling : $p(x)$ or $p(y)$; Conditional sampling : $p(x|y)$ or $p(y|x)$; Joint sampling : $p(x, y)$. **The key idea is to have a diffusion step per modality !**

Algorithm 1 Training

repeat
  $\boldsymbol{x}_0, \boldsymbol{y}_0 \sim q(\boldsymbol{x}_0, \boldsymbol{y}_0)$
  $t^x, t^y \sim \text{Uniform}(\{1, 2, \dots, T\})$
  $\boldsymbol{\epsilon}^x, \boldsymbol{\epsilon}^y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  Let $\boldsymbol{x}_{t^x} = \sqrt{\overline{\alpha}_{t^x}}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_{t^x}}\boldsymbol{\epsilon}^x$
  Let $\boldsymbol{y}_{t^y} = \sqrt{\overline{\alpha}_{t^y}}\boldsymbol{y}_0 + \sqrt{1-\overline{\alpha}_{t^y}}\boldsymbol{\epsilon}^y$
  Take gradient step on
    $\nabla_{\boldsymbol{\theta}}\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t^x}, \boldsymbol{y}_{t^y}, t^x, t^y) - [\boldsymbol{\epsilon}^x, \boldsymbol{\epsilon}^y]\|_2^2$
until converged

Algorithm 2 Unconditional sampling of $\boldsymbol{x}_0$ (similar for unconditional sampling of $\boldsymbol{y}_0$)

$\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T, \dots, 1$ **do**
  $\boldsymbol{z}^x \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z}^x = \mathbf{0}$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}^x_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{\epsilon}, t, T)\right) + \sigma_t \boldsymbol{z}^x$
end for
return $\boldsymbol{x}_0$

Algorithm 3 Sampling of $\boldsymbol{x}_0$ conditioned on $\boldsymbol{y}_0$ (similar for sampling of $\boldsymbol{y}_0$ conditioned on $\boldsymbol{x}_0$)

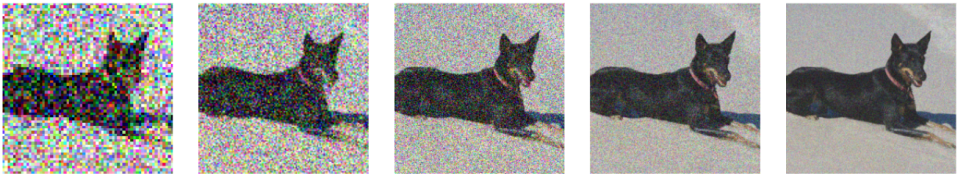
$\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T, \dots, 1$ **do**
  $\boldsymbol{z}^x \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z}^x = \mathbf{0}$
  $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}^x_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t, 0)\right) + \sigma_t \boldsymbol{z}^x$
end for
return $\boldsymbol{x}_0$

Algorithm 4 Joint sampling of $\boldsymbol{x}_0, \boldsymbol{y}_0$

$\boldsymbol{x}_T, \boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T, \dots, 1$ **do**
  $\boldsymbol{z}^x, \boldsymbol{z}^y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z}^x, \boldsymbol{z}^y = \mathbf{0}$
  $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}^x_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_t, t, t)\right) + \sigma_t \boldsymbol{z}^x$
  $\boldsymbol{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}^y_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_t, t, t)\right) + \sigma_t \boldsymbol{z}^y$
end for
return $\boldsymbol{x}_0, \boldsymbol{y}_0$

Figure 6: From [Bao et al. 2023a], 2023

We now have the background elements to have a multimodal/joint diffusion model with perceiver-IO!
But multimodality in Remote Sensing can apply multi resolution. Some previous work [T. Chen 2023 Falck et al. 2025]demonstrated that we need to be careful with Diffusion model and multi resolution.



(a) 64×64 (b) 128×128 (c) 256×256 (d) 512×512 (e) 1024×1024

Figure 2: Noised images ($\boldsymbol{x}_t = \sqrt{\gamma}\boldsymbol{x}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}$) with the same noise level ($\gamma = 0.7$). We see that higher resolution natural images tend to exhibit higher degree of redundancy in (nearby) pixels, therefore less information is destroyed with the same level of independent noise.

Figure 7: From T. Chen 2023

Limitation of existing work

- Multimodal Models have an encoder per modality

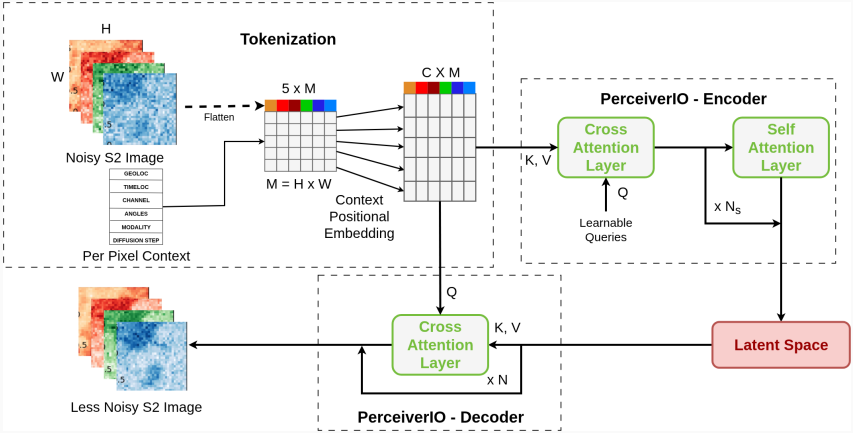
- Most model only generate Sentinel 2

- Conditional model are text based

- Multimodal Model are not multiresolution

- Multimodal Model are limited to S2 / S1

Diffusion with Perceiver-IO : Model



Diffusion with Perceiver-IO : Training that gave satisfying results

- **Dataset :** 32x32 samples of Sentinel 2A with Bands : [B02, B03, B04, B08]. Train : 5000 samples / Val : 500 samples
- **Context Content :**
  - Coordinate of pixel in sequence (512 features)
  - Channel (16 features)
  - Diffusion Step (512 features)

Figure 8: Training Areas of Interest (AOI)

Diffusion with Perceiver-IO : A way to witness a per step diffusion performance

Measuring Peak Signal to Noise Ratio during forward phase.
For each diffusion step we compute x_t and estimate x_0 with our diffusion model.

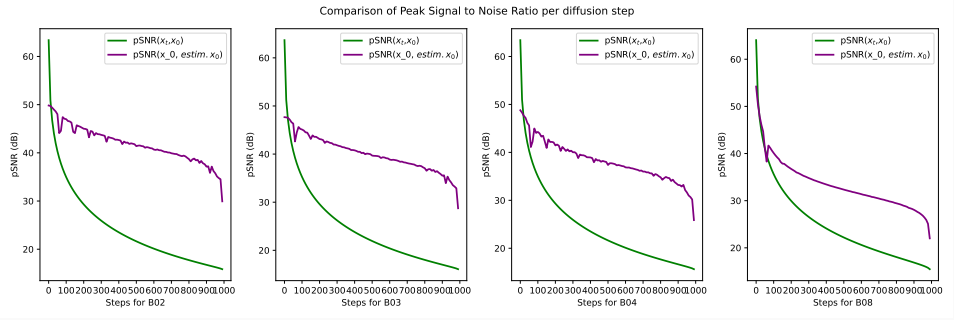


Figure 9: PSNR per diffusion step averaged over 50 samples

The goal is to evaluate the capacity of the diffusion model to denoise the data per diffusion step.

Diffusion with Perceiver-IO : Unconditional Setting

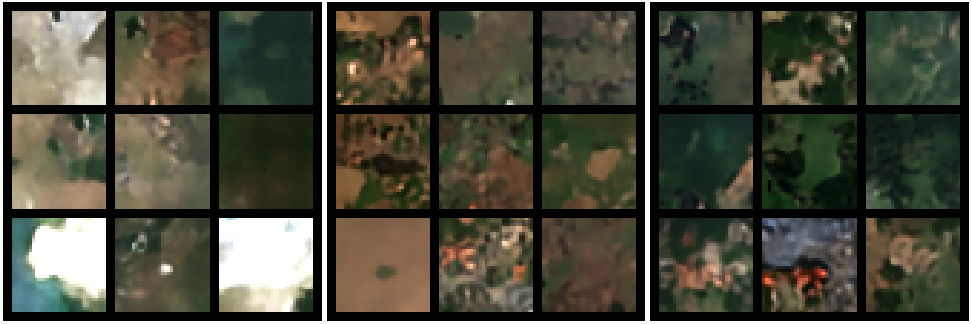


Figure 10: Unconditionnal Generation of Sentinel2 RGB images

Diffusion with Perceiver-IO : Unconditional Setting

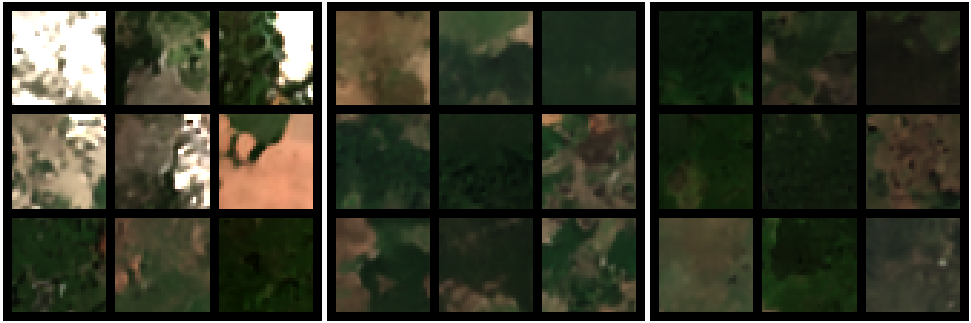


Figure 11: Unconditionnal Generation of Sentinel2 RGB images

Diffusion with Perceiver-IO : Unconditional Setting - Distribution

What about the distribution?

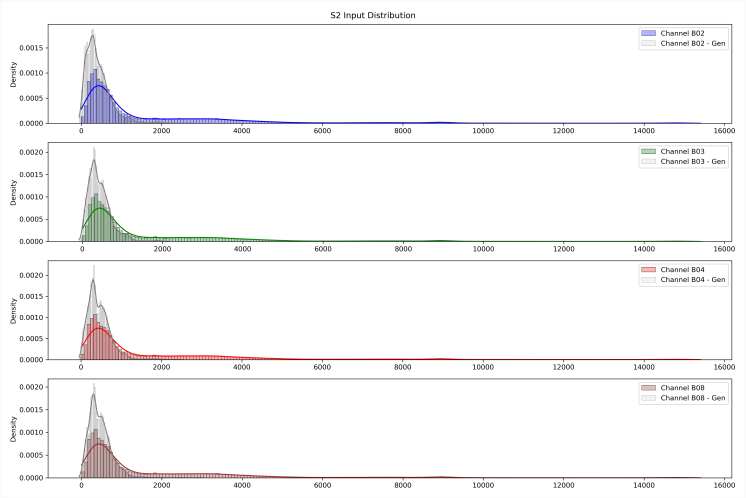


Figure 12: Distribution per band of 100 train samples (colored) and 100 generated samples (grey)

Diffusion with Perceiver-IO : Unconditional Setting - Closest Sample from training set

Memorization or Generalization ?

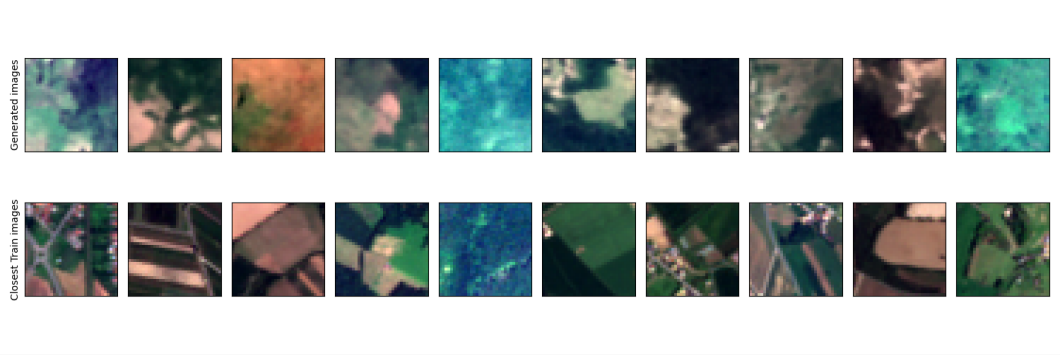


Figure 13: Closest images of train set for different generated images. Distance measured with cosine similarity. Code used is taken from Kadkhodaie et al. 2024

Diffusion with Perceiver-IO : Unconditional Setting - Power Spectral Density (PSD)

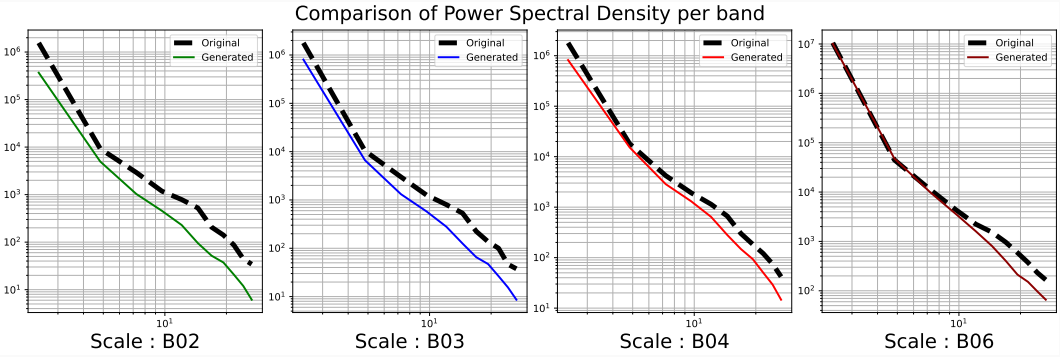


Figure 14: Comparison of the Averaged Power Spectral Density (PSD) for 150 generated samples and 150 training samples.

The PSD of generated samples is lower than the PSD training samples. (It means that the samples are more blurry.) Still, the majority of the scales are well reproduced!

Diffusion with Perceiver-IO : Conclusions

We are now sure about :

- The way of encoding the diffusion steps (Positional encoding in the context of the perceiver-IO)

- The fact that the perceiver-IO is able to generate images with diffusion process.

- It seems that, above all, able to catch the diversity of the data and is able to reproduce it.

- The setup from [Bao et al. 2023b] works with our Perceiver-IO setting and allows us to have a multimodal/joint Diffusion model.

PhD Perspectives

1 - **Train a Multimodal/Joint Diffusion Model and explore conditioning on Auxiliary data (ex: Meteo)** [Leonhardt, Gall, and Roscher 2025,Goktepe et al. 2025]

2 - **Explore Scalability of Diffusion Models** [Karras, Aittala, Aila, et al. n.d.,Lu, Zhou, J. Chen, et al. n.d.,Lu, Zhou, Bao, et al. 2025]

3 - **Compare performance with current State of the Art models and architecture**

4 - **Investigate performance with Flow Matching approach** [Le Bellier and Audebert 2026]

5 - **Use-case RELEO: Uncertainty Quantification of Essential Variables**

Thanks for your attention!

Figure 15: From pinterest : 101 Funny Thank You Memes to Say Thanks for a Job Well Done

Appendix

Diffusion with Perceiver-IO : Conditional Setting - Ongoing work

- With the same training setting and the same model architecture, we train the model to generate [B02, B03, B04, B08] in **Any-to-any** setting.

- The model can be either unconditional or conditional.

- We now have 14 possibilities of generation depending on the conditioning bands. One can condition on either a single band or on two or three bands the generation.

- Currently I am using different metrics to evaluate the conditioning generation (CRPS, QQ-plot, Pair-plot KDE)

Diffusion with Perceiver-IO : Conditional Setting - Early Results

We sample one band conditioned on the others