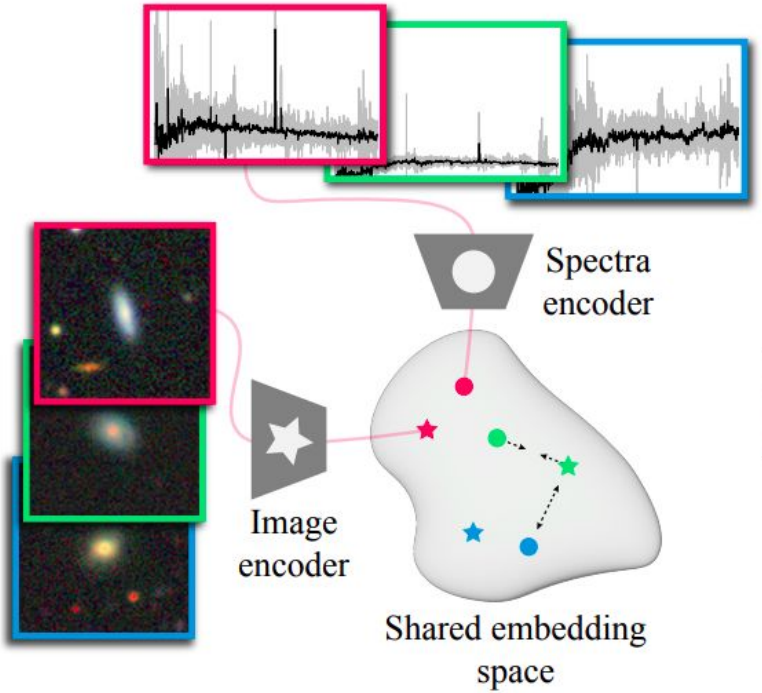
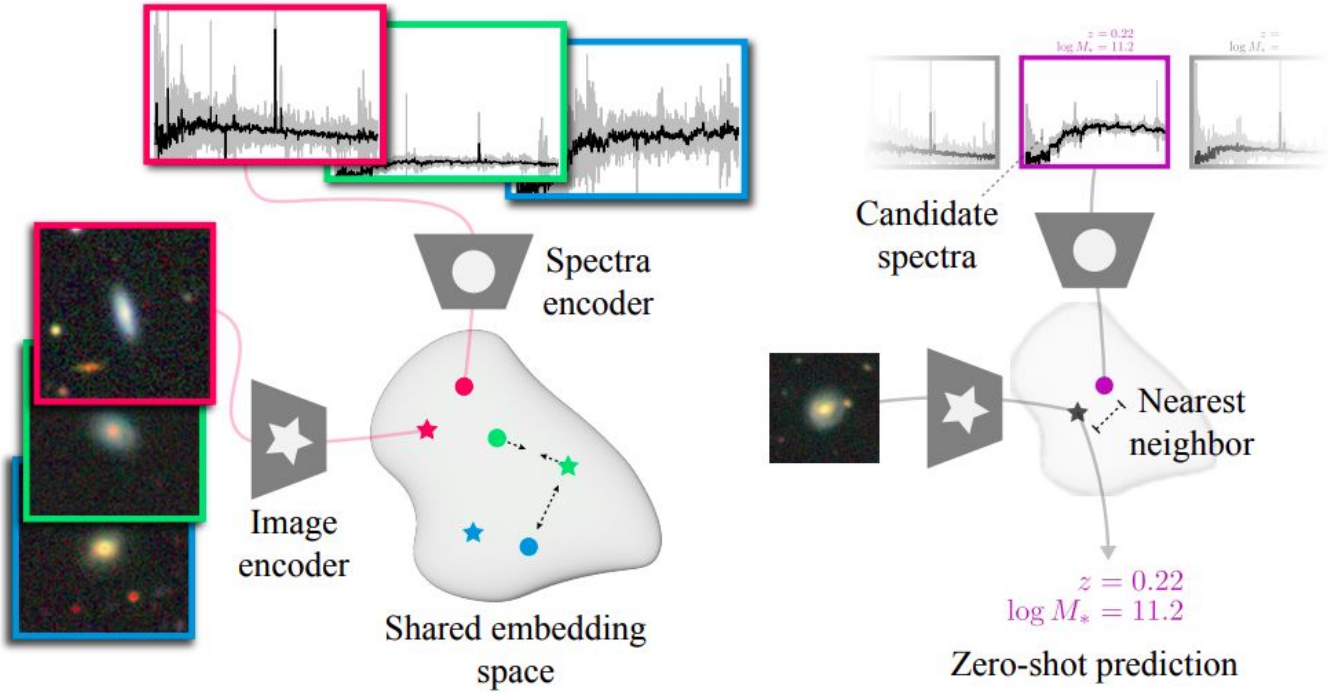
AstroCLIP: A Cross-Modal Foundation Model for Galaxies

Liam Parker,[1] ⋆ † Francois Lanusse,[1,3] Siavash Golkar,[1] Leopoldo Sarra,[1] Miles Cranmer,[4]
Alberto Bietti,[1] Michael Eickenberg,[1] Geraud Krawezik,[1] Michael McCabe,[1,5] Rudy Morel,[1] Ruben Ohana,[1]
Mariel Pettee,[1,6] Bruno Régaldo-Saint Blancard,[1] Kyunghyun Cho,[7,8,9] Shirley Ho[1,7,10] and
The Polymathic AI Collaboration

AstroCLIP: A Cross-Modal Foundation Model for Galaxies



AstroCLIP: A Cross-Modal Foundation Model for Galaxies



AstroCLIP: A Cross-Modal Foundation Model for Galaxies

Liam Parker,[1] ⋆ † Francois Lanusse,[1,3] Siavash Golkar,[1] Leopoldo Sarra,[1] Miles Cranmer,[4]
Alberto Bietti,[1] Michael Eickenberg,[1] Geraud Krawezik,[1] Michael McCabe,[1,5] Rudy Morel,[1] Ruben Ohana,[1]
Mariel Pettee,[1,6] Bruno Régaldo-Saint Blancard,[1] Kyunghyun Cho,[7,8,9] Shirley Ho[1,7,10] and
The Polymathic AI Collaboration

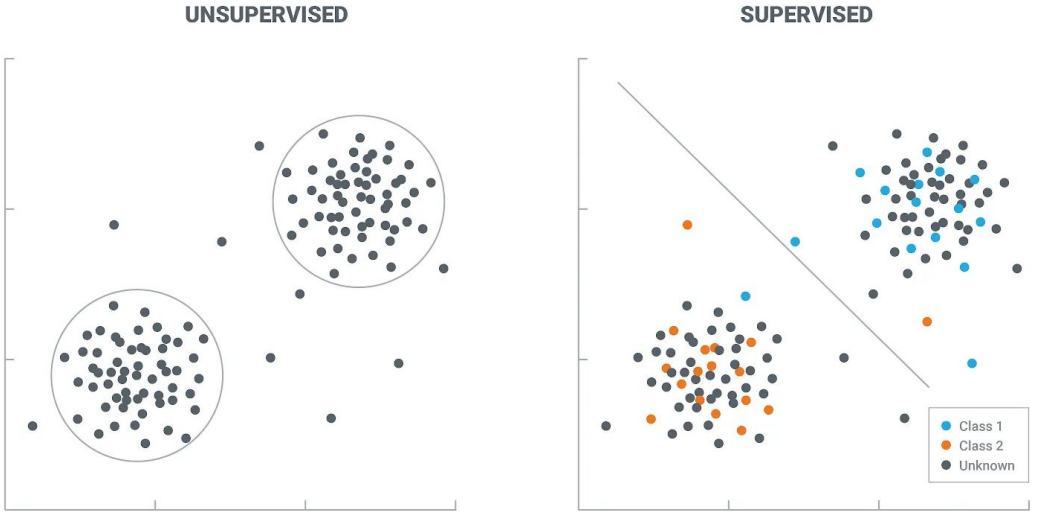
ABSTRACT

We present AstroCLIP, a single, versatile model that can embed both galaxy images and spectra into a shared, physically meaningful latent space. These embeddings can then be used - without any model fine-tuning - for a variety of downstream tasks including (1) accurate in-modality and cross-modality semantic similarity search, (2) photometric redshift estimation, (3) galaxy property estimation from both images and spectra, and (4) morphology classification. Our approach to implementing AstroCLIP consists of two parts. First, we embed galaxy images and spectra separately by pretraining separate transformer-based image and spectrum encoders in self-supervised settings. We then align the encoders using a contrastive loss. We apply our method to spectra from the Dark Energy Spectroscopic Instrument and images from its corresponding Legacy Imaging Survey. Overall, we find remarkable performance on all downstream tasks, even relative to supervised baselines. For example, for a task like photometric redshift prediction, we find similar performance to a specifically-trained ResNet18, and for additional tasks like physical property estimation (stellar mass, age, metallicity, and sSFR), we beat this supervised baseline by 19% in terms of R^2. We also compare our results to a state-of-the-art self-supervised single-modal model for galaxy images, and find that our approach outperforms this benchmark by roughly a factor of two on photometric redshift estimation and physical property prediction in terms of R^2, while remaining roughly in-line in terms of morphology classification. Ultimately, our approach represents the first cross-modal self-supervised model for galaxies, and the first self-supervised transformer-based architectures for galaxy images and spectra.

1 Introduction

- **Supervised** methods

- **Unsupervised** methods



1 Introduction

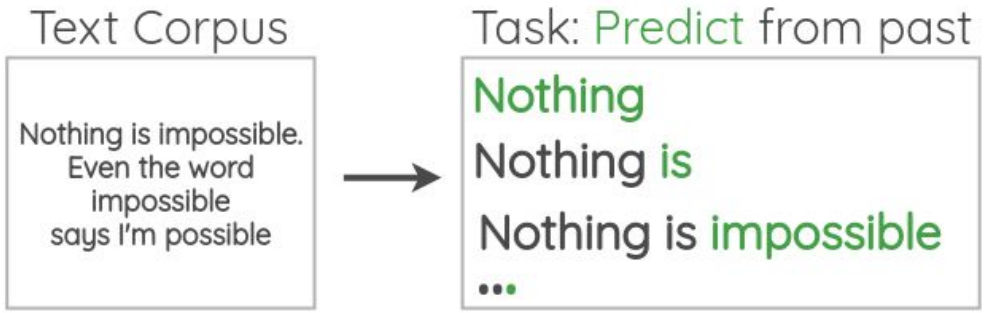
- **Supervised** methods

- **Unsupervised** methods

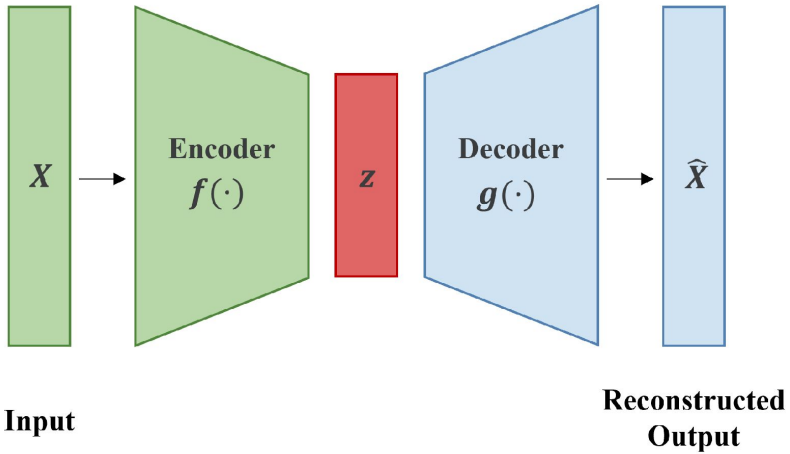
Recently, a new line of inquiry has explored **self-supervised learning (SSL)** as an alternative. These approaches learn high-quality embeddings i.e. low-dimensional representations of the objects that preserve their important physical information - in the absence of labeled training data. These embeddings can then be used for a variety of downstream tasks, eliminating the need to retrain bespoke supervised models from scratch for each new dataset or new task.

1 Introduction

- **Supervised** methods
- **Unsupervised** methods



Recently, a new line of inquiry has explored **self-supervised learning (SSL)** as an alternative. These approaches learn high-quality embeddings i.e. low-dimensional representations of the objects that preserve their important physical information - in the absence of labeled training data. These embeddings can then be used for a variety of downstream tasks, eliminating the need to retrain bespoke supervised models from scratch for each new dataset or new task.



1 Introduction

The main contributions of our work are:

- We develop the first self-supervised transformer-based models for galaxy spectra and images.
- We apply a cross-modal training regime to align the pre-trained image and spectrum encoders around shared physical semantics, creating a unified latent space for spectra and images.
- We empirically demonstrate that our cross-modal embeddings capture core physical properties of the underlying galaxies. This enables, with only minimal downstream processing, AstroCLIP to be used for:
  - In-modal and cross-modal galaxy similarity searches.
  - Photometric redshift estimation
  - Galaxy property estimation from images
  - Galaxy property estimation from spectra
  - Galaxy morphology classification from images.

1 Introduction

Code for our models, training and testing kit is available online here.

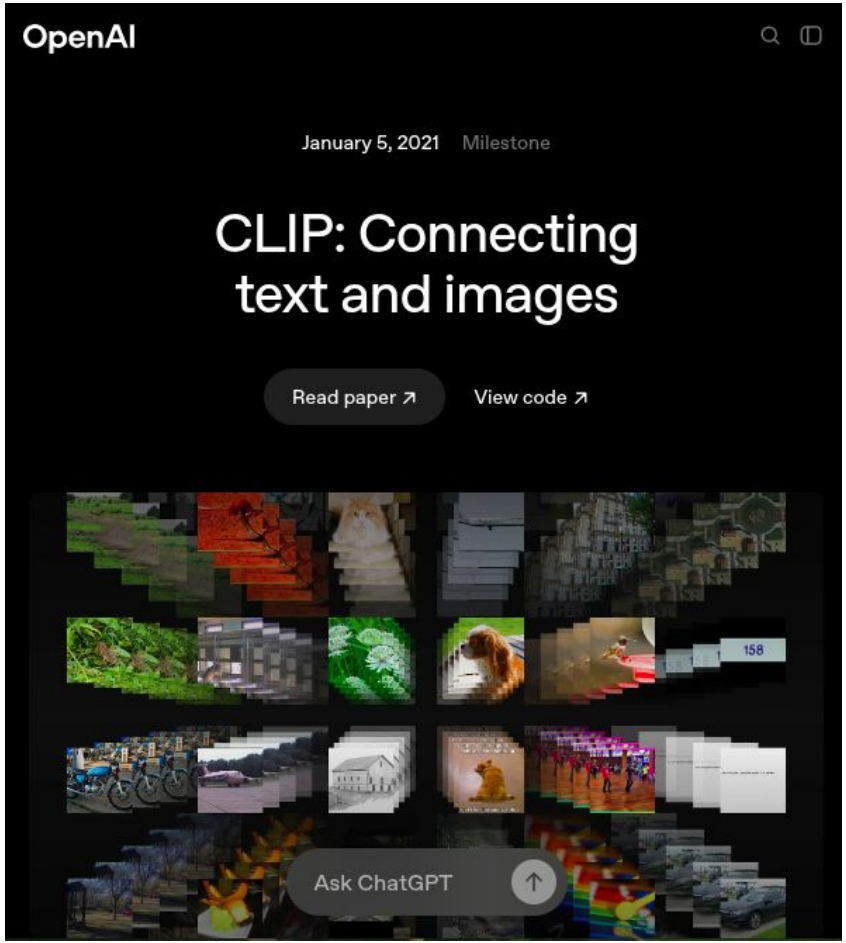
Our paper is organized as follows. In section 2, we provide background on self-supervised learning, as well as on the particular SSL objectives used in the present paper. In section 3, we describe the specifics of our AstroCLIP implementation. In section 4, we provide the data sets that we use to train our models and in section 5, we outline the training process of our models. In section 6, we present our results on in-modal and cross-modal similarity searches, photometric redshift estimation, galaxy property prediction, and morphology classification. Finally, we discuss our results and further extensions of our paper in section 7.

2 Self-Supervised Learning

In self-supervised learning (SSL), the objective is to train a model to learn to extract rich, low-dimensional representations from data without the need for any labels. This is typically achieved by training the model to perform some contrived surrogate task on the input data. In recent years, a variety of such surrogate tasks have been developed. One common example of such a task in NLP is to predict the next word in a sentence given the previous words; this is typically called autoregressive prediction (Radford et al. 2019). Many other such objectives have been developed, including masked reconstruction (Devlin et al. 2018; He et al. 2021), self-distillation (Fang et al. 2021), and contrastive learning (Chen et al. 2020; Radford et al.

2 Self-Supervised Learning

2.1 Cross-Modal Contrastive Techniques



2 Self-Supervised Learning

2.1 Cross-Modal Contrastive Techniques

$$\mathcal{L}_{InfoNCE}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp(S_C(\mathbf{x}_i, \mathbf{y}_i)/\tau)}{\sum_j^K \exp(S_C(\mathbf{x}_i, \mathbf{y}_j)/\tau)}. \quad (1)$$

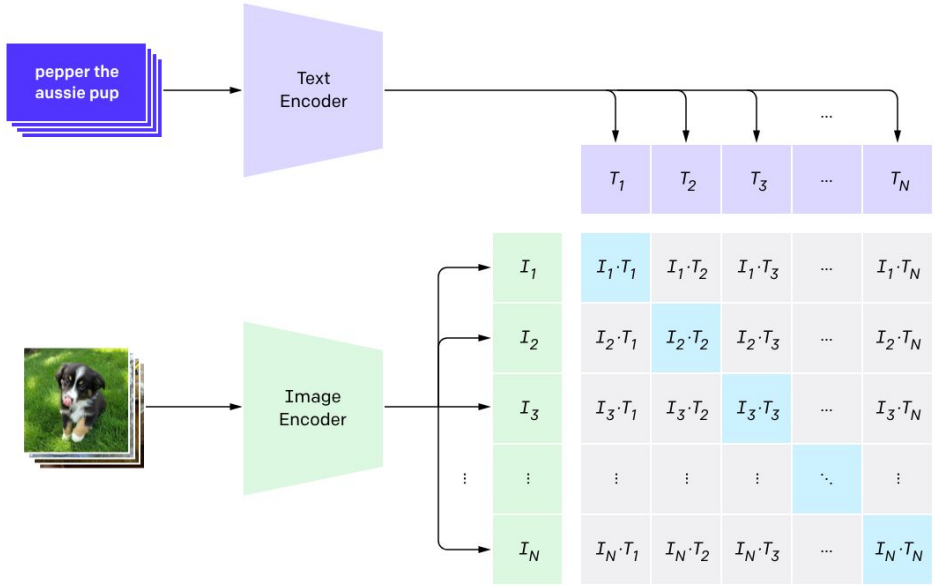
Here, $\tau > 0$ represents a smoothing parameter (sometimes referred to as temperature) and j represent the indices of negative examples not associated with the object i.

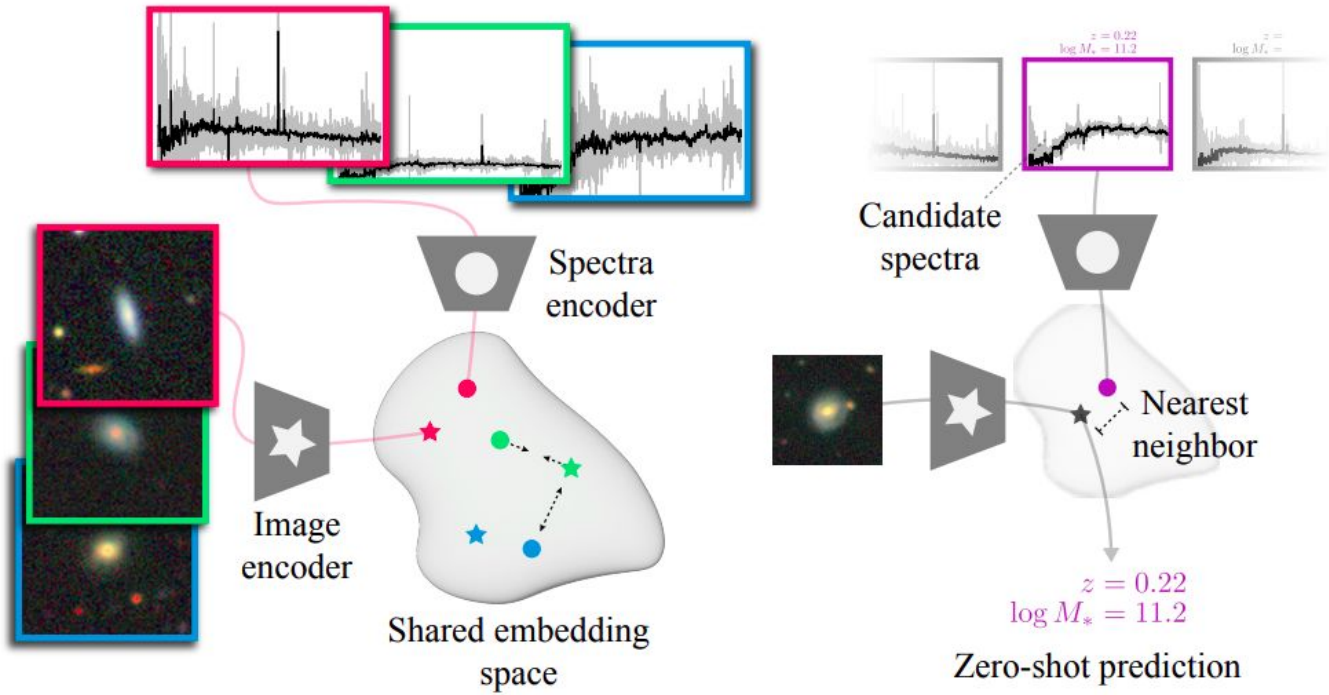
Additionally, a choice of similarity metric, S_C, must be specified to determine the similarity between representations in the embedding space. In CLIP, the cosine similarity between two points in the embedding space is used, such that

$$S_C(\mathbf{x}_i, \mathbf{y}_j) = \frac{(\mathbf{x}_i)^T \mathbf{y}_j}{\|\mathbf{x}_i\|_2^2 \|\mathbf{y}_j\|_2^2}. \quad (2)$$

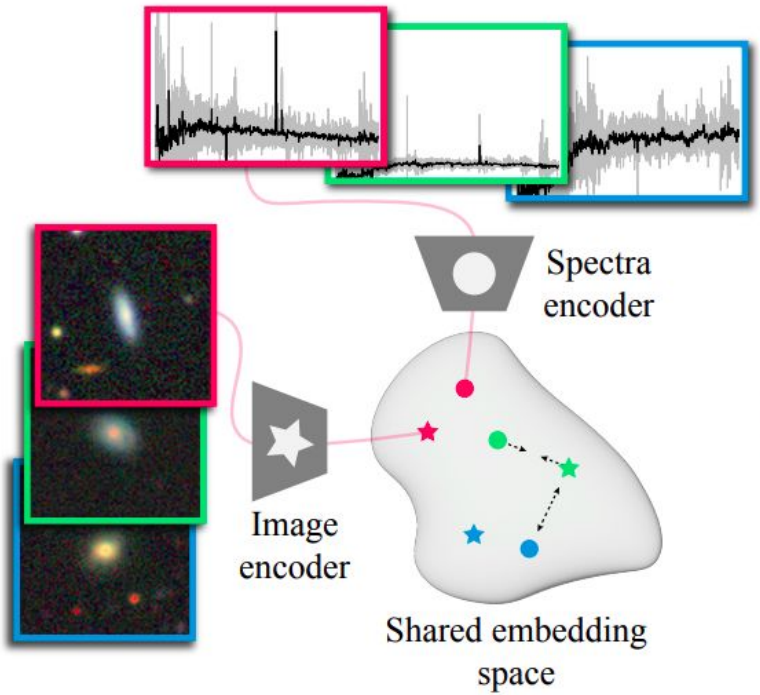
Intuitively, the InfoNCE objective works by bringing together points in the embedding space that correspond to the same underlying physical object and pushing points in the embedding space away from each other if they correspond to different underlying physical objects. Be-

1. Contrastive pre-training

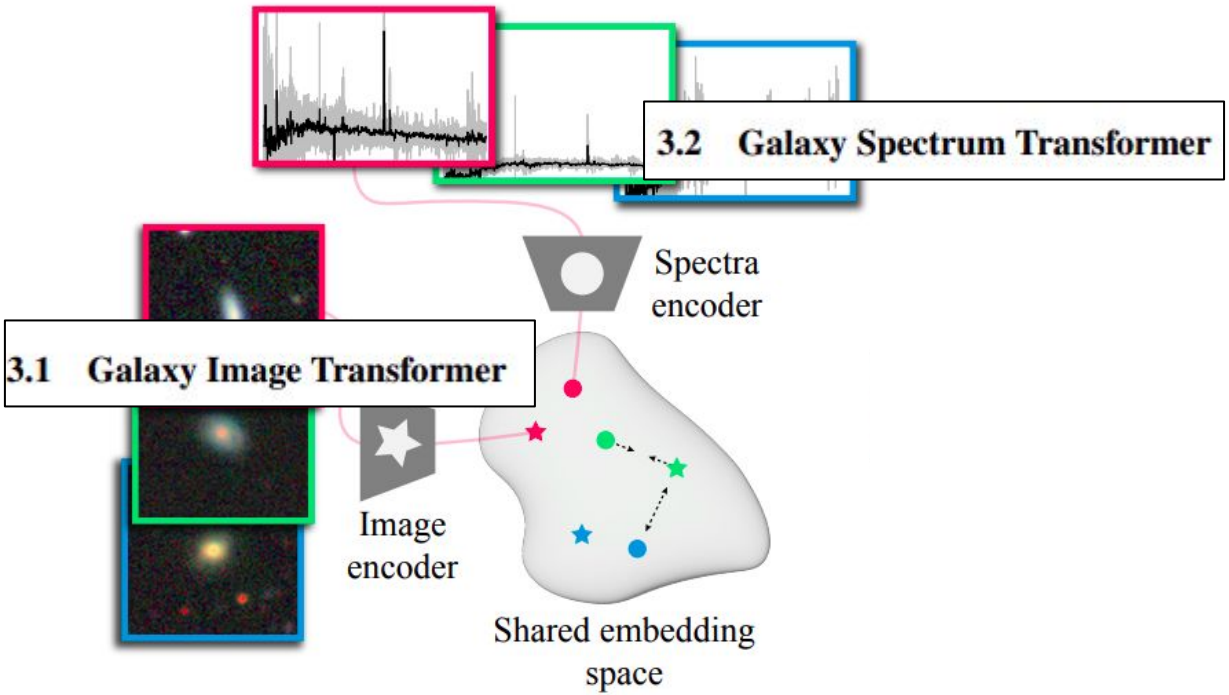



Spectra encoder
Image encoder
Shared embedding space
Candidate spectra
Nearest neighbor
$z = 0.22$
$\log M_* = 11.2$
Zero-shot prediction

3 AstroCLIP Model Implementation



3 AstroCLIP Model Implementation



Flow Matching

Creating noise from data is easy; creating data from noise is generative modeling.

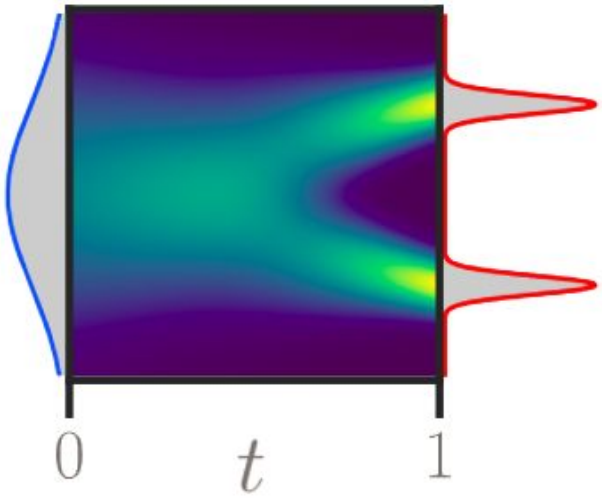
Song et al.

Probability Path

$$P(X|t)$$

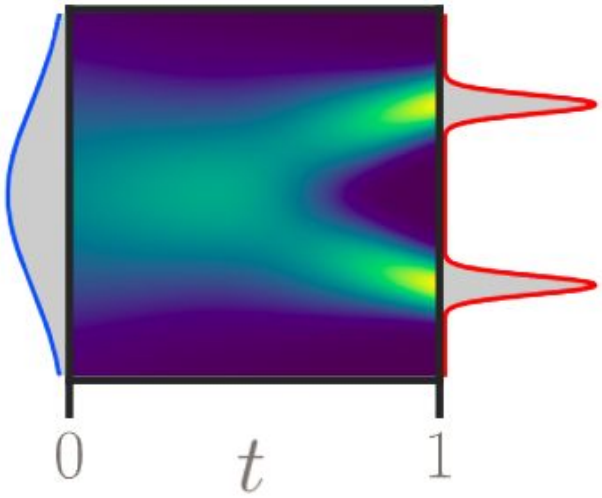
Creating noise from data is easy; creating data from noise is generative modeling.

Song et al.

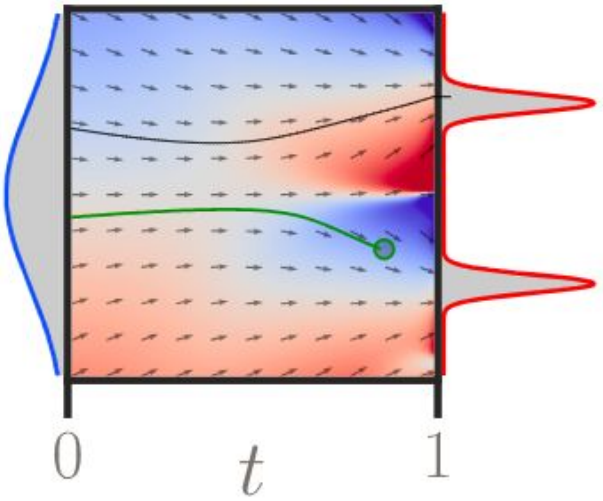


Probability Path

$$P(X|t)$$



“Velocity field”



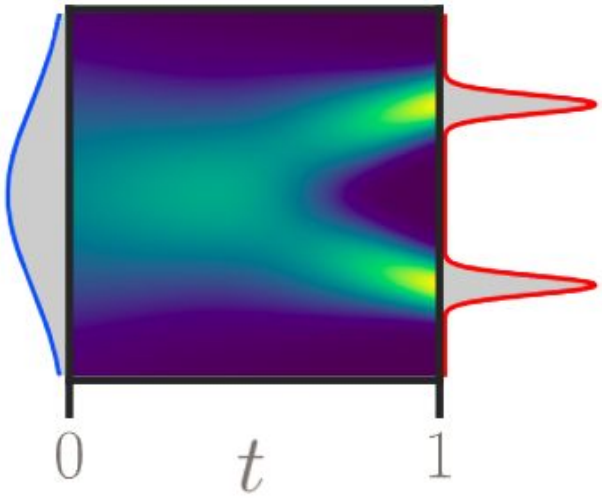
Probability Path

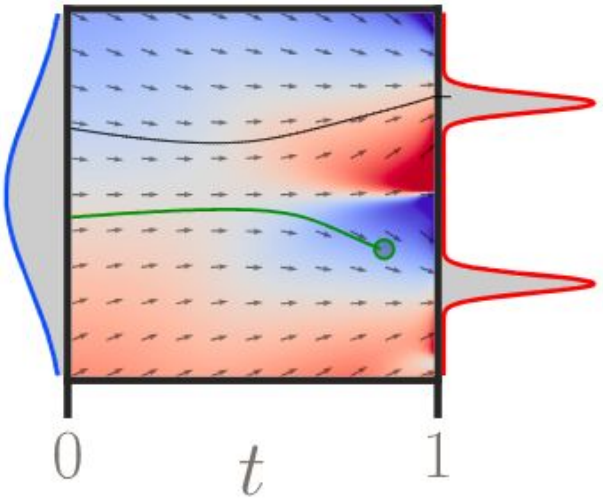
$$P(X|t)$$

“Velocity field”

$$V(X,t)$$

$$\mathbb{R}^{dim(X,t)} \rightarrow \mathbb{R}^{dim(X)}$$



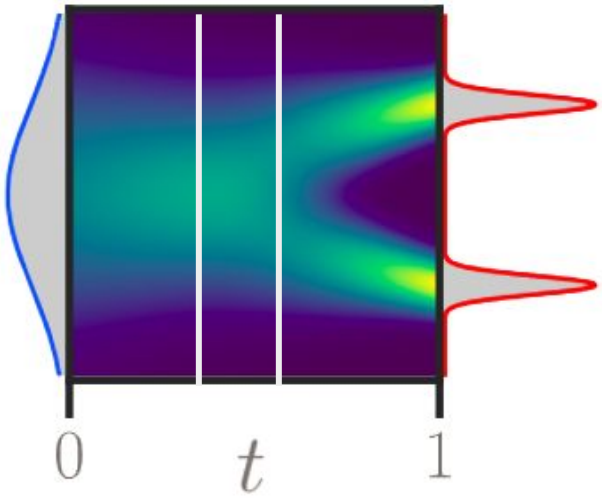


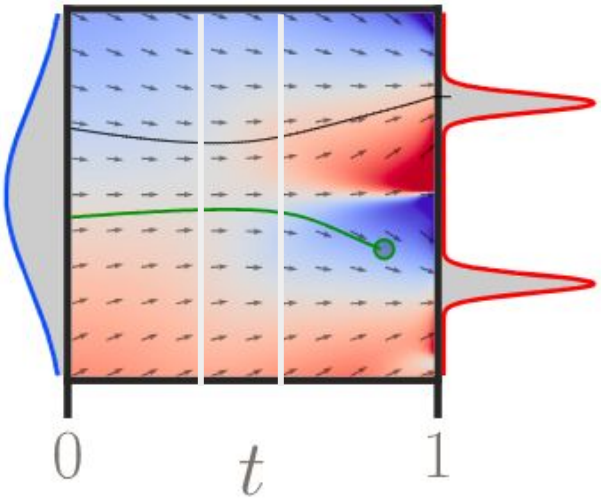
Probability Path

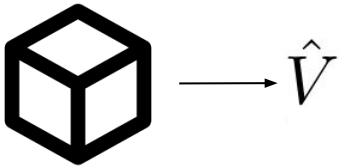
"Velocity field"

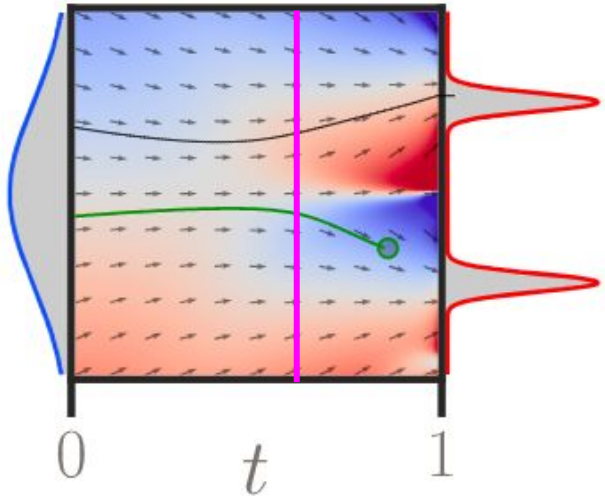
$$x_0 \sim P(X|t_0), \quad x_1 = x_0 + V(x_0, t_0)$$

$$\implies x_1 \sim P(X|t_1)$$



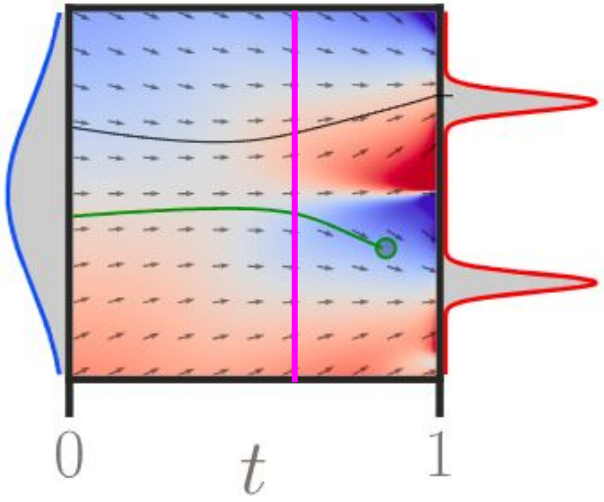



$\hat{V}$

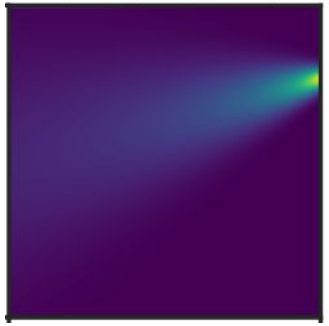

0
t
1

$Z \sim \text{data}$

$\longrightarrow \hat{V}$

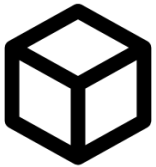


$p(x, t|z = z^{(1)})$

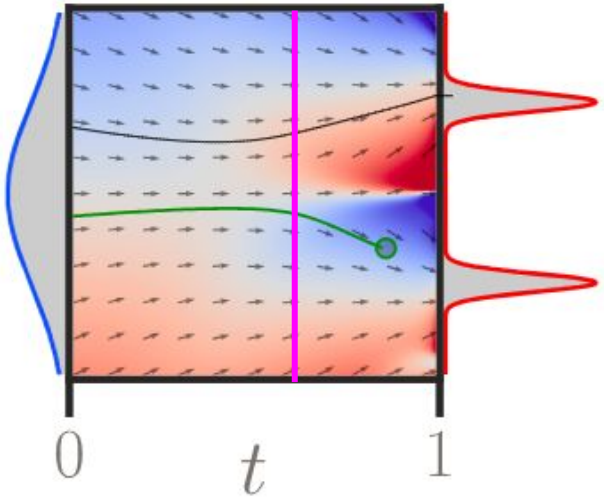


$Z \sim \text{data}$

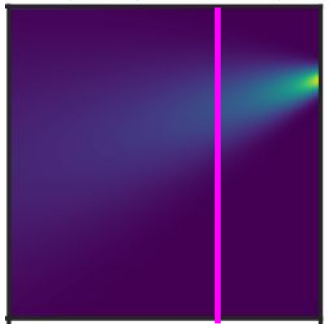
$t \sim U(0,1)$



$\hat{V}$



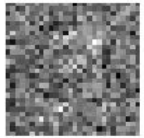
$p(x, t|z = z^{(1)})$

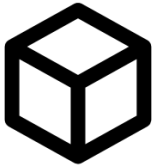


$Z \sim \text{data}$

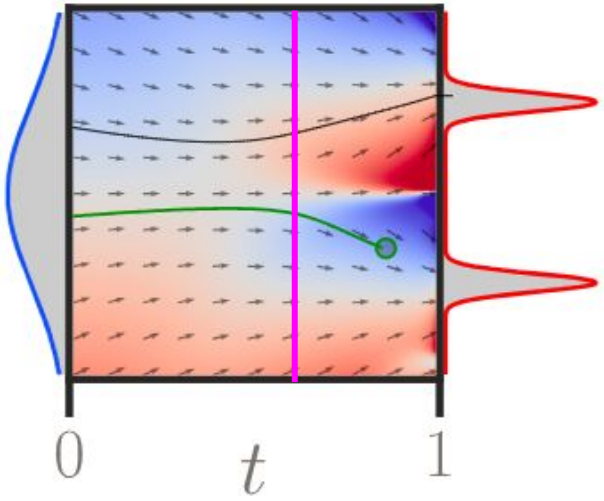
$t \sim U(0,1)$

$x \sim P(X|t,Z)$

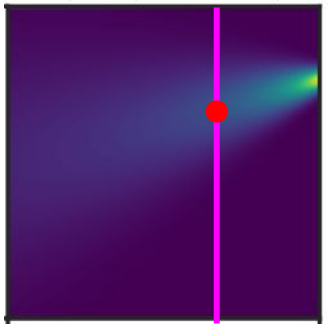




$\hat{V}$

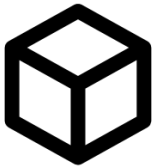


$p(x, t|z = z^{(1)})$

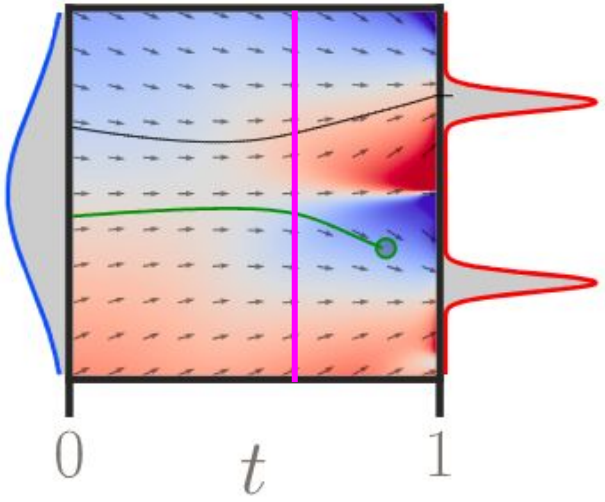


$Z \sim \text{data}$

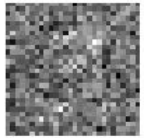
$t \sim U(0,1)$



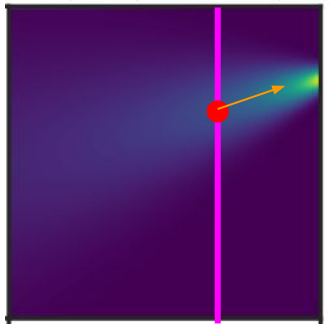
$\hat{V} - V(X,t|Z)$



$x \sim P(X|t,Z)$



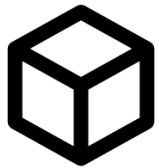
$p(x,t|z=z^{(1)})$



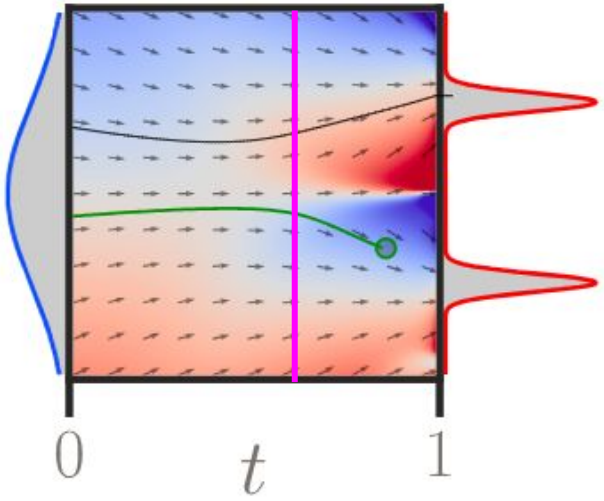
$Z \sim \text{data}$

$t \sim U(0,1)$

$x \sim P(X|t,Z)$



$\hat{V} - V(X,t|Z)$

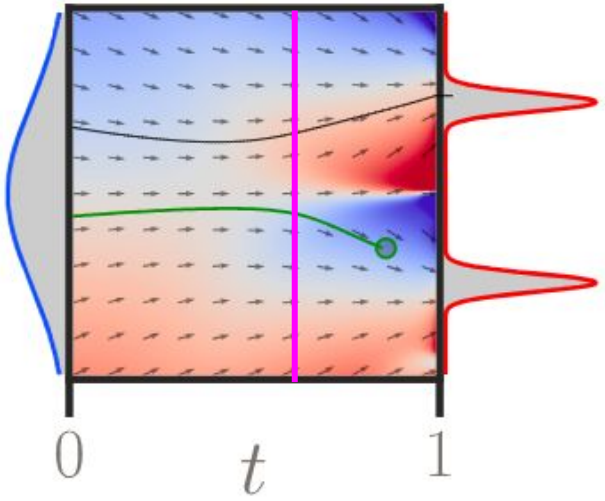


$Z \sim \text{data}$

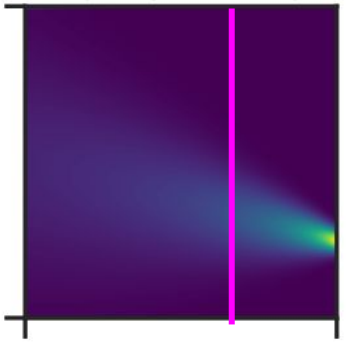
$t \sim U(0,1)$

$x \sim P(X|t,Z)$

$\hat{V} - V(X,t|Z)$

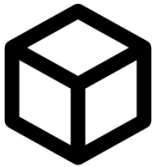


$p(x,t|z=z^{(2)})$

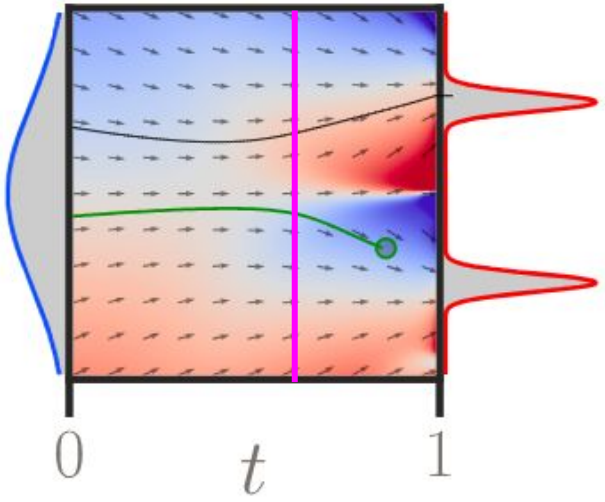


$Z \sim \text{data}$

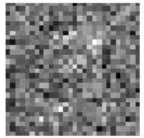
$t \sim U(0,1)$



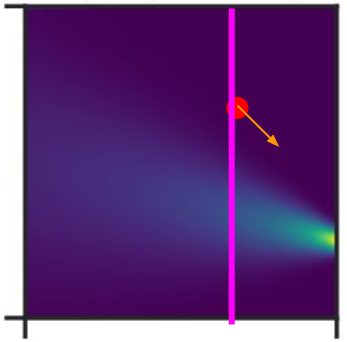
$\hat{V} - V(X,t|Z)$

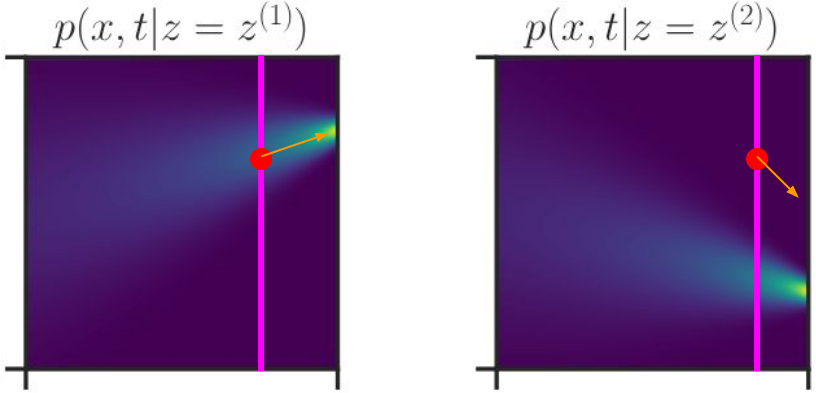


$x \sim P(X|t,Z)$

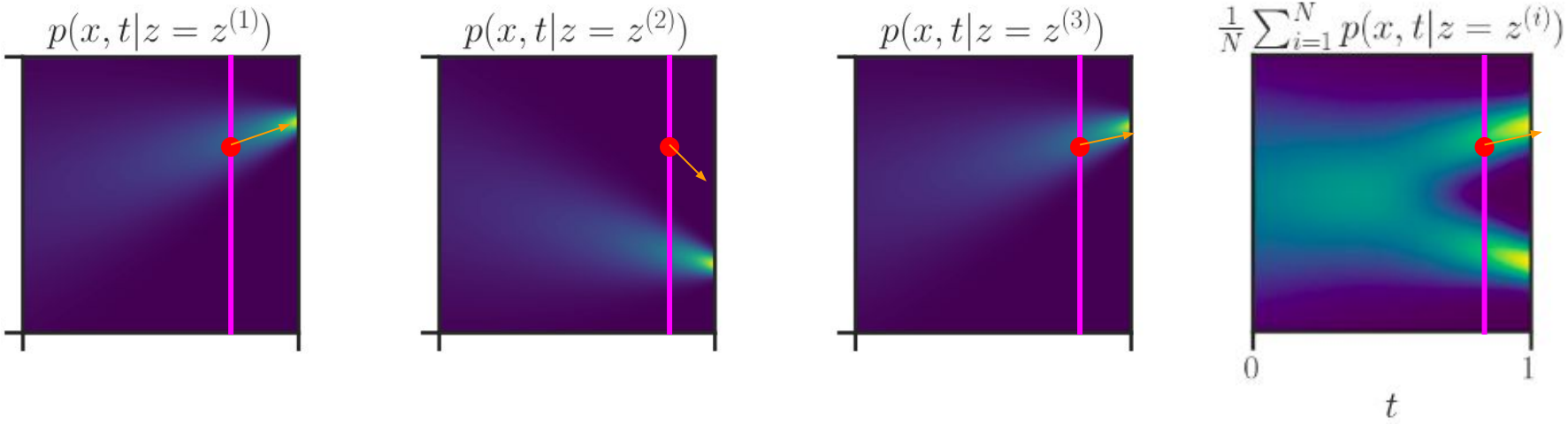


$p(x,t|z=z^{(2)})$

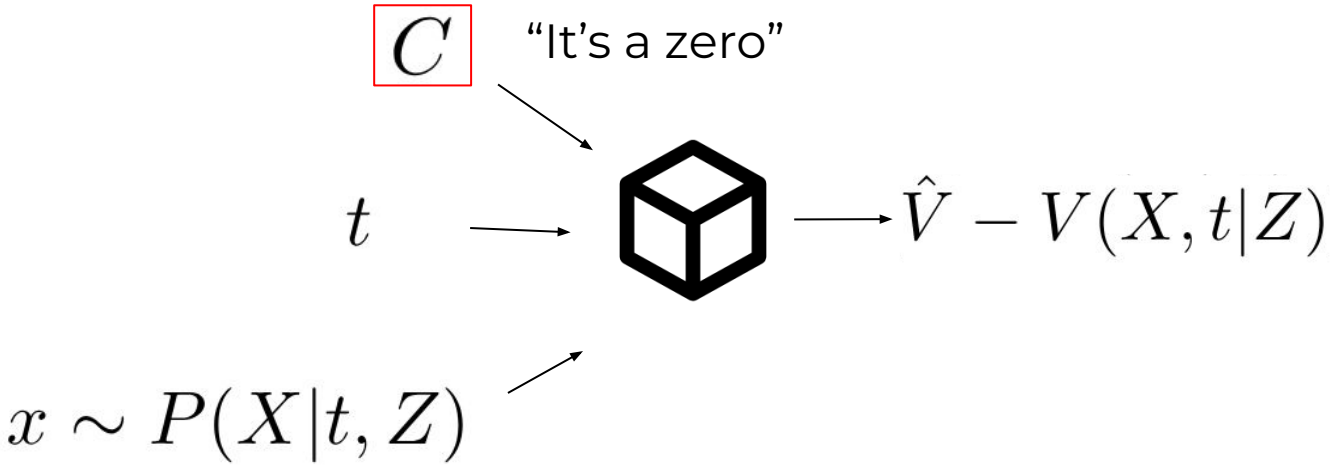


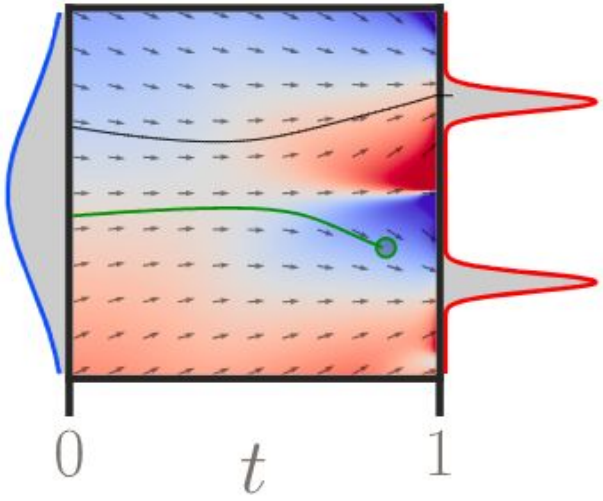


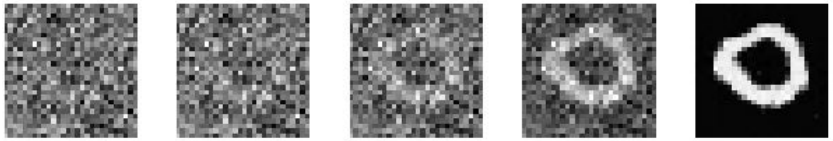
$$V(X,t) = E_z[V(X,t|Z)]$$



$$V(X,t) = E_z[V(X,t|Z)]$$


C
"It's a zero"
t
$\hat{V} - V(X, t|Z)$
$x \sim P(X|t, Z)$

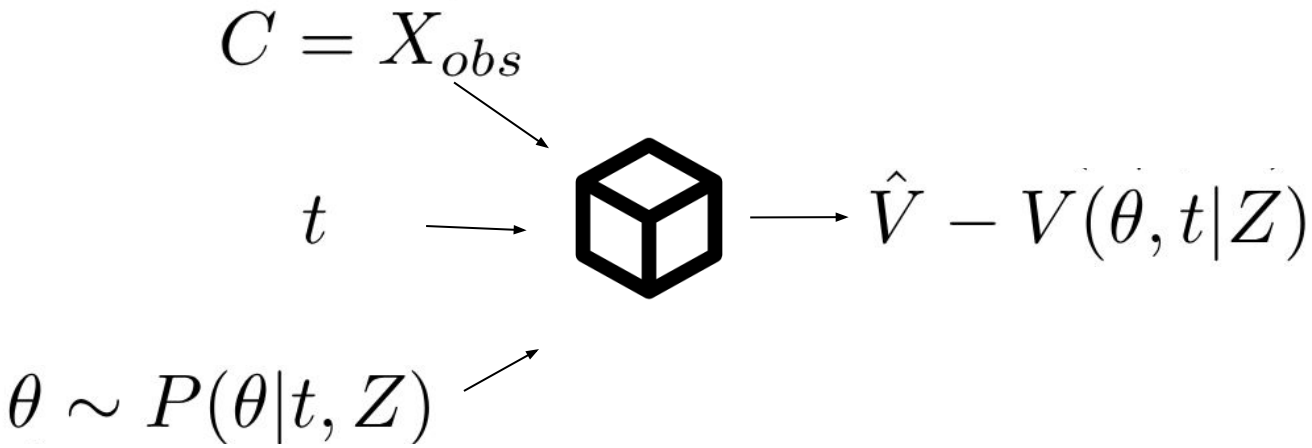

0
t
1



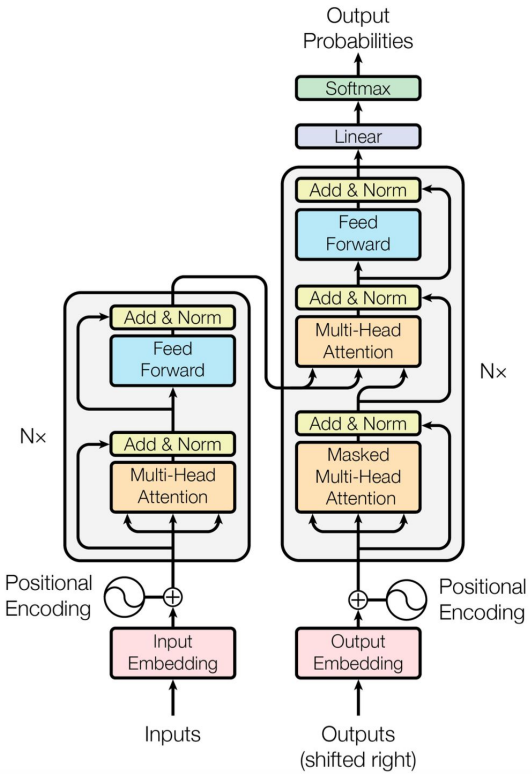
Simulation Based Inference

$$\theta \longrightarrow X_{obs}$$

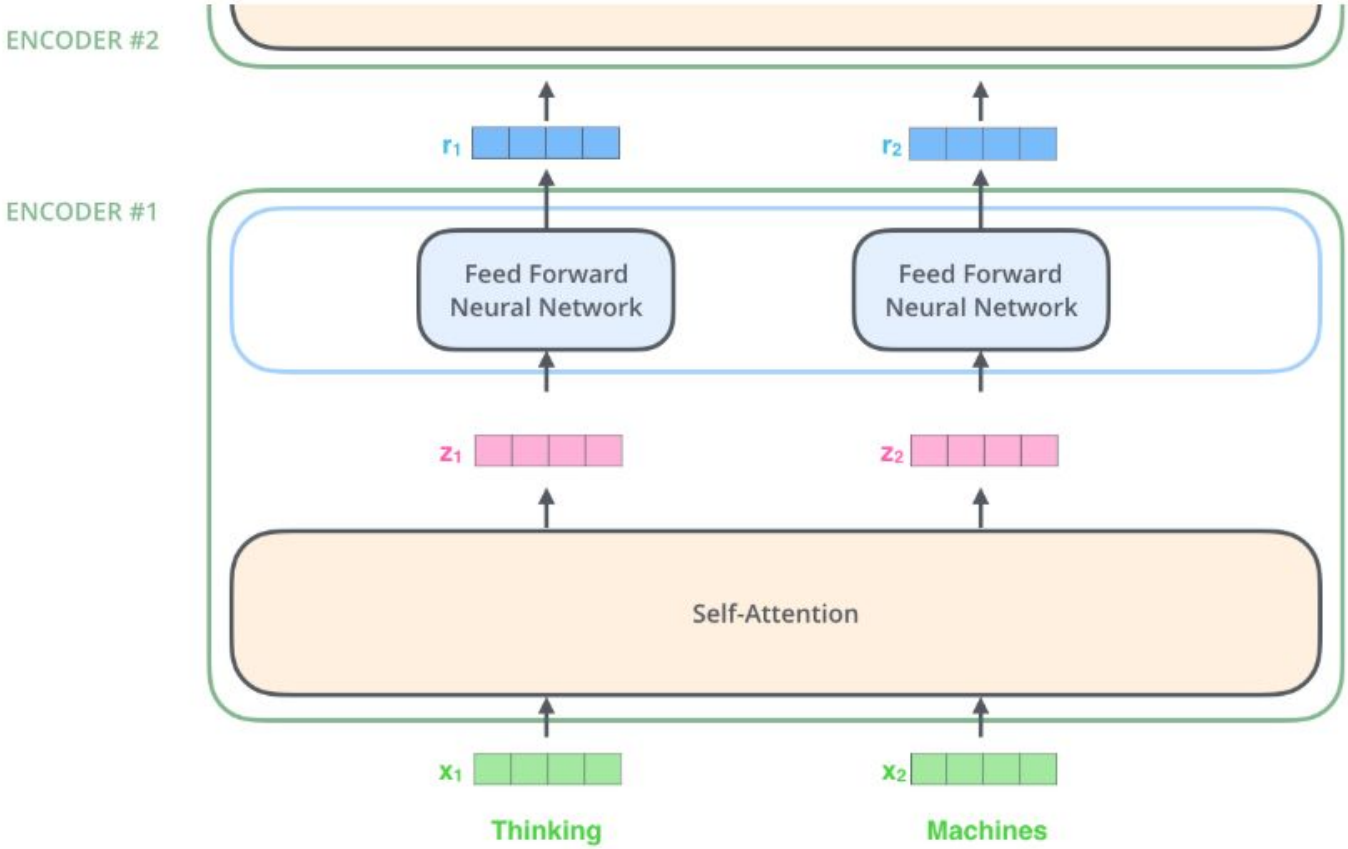
Simulation Based Inference



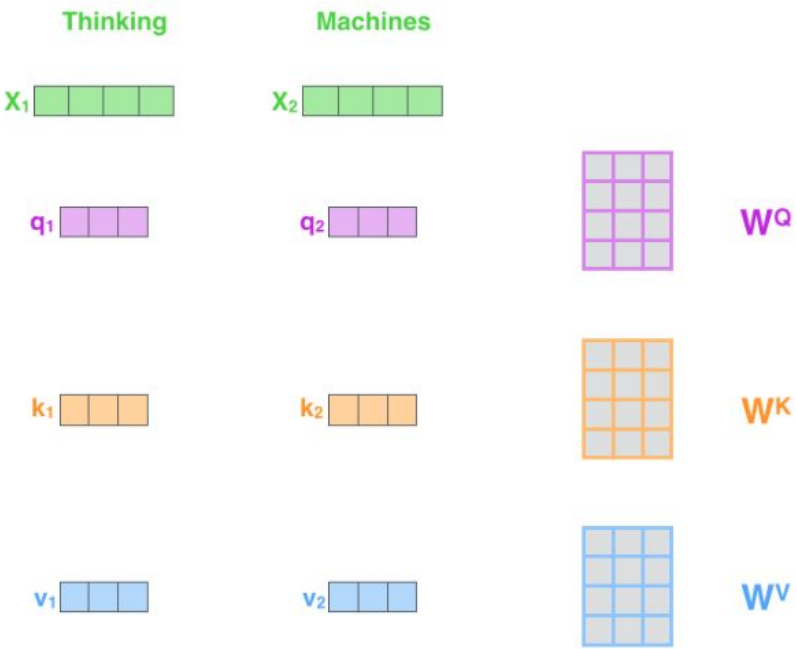
Transformers



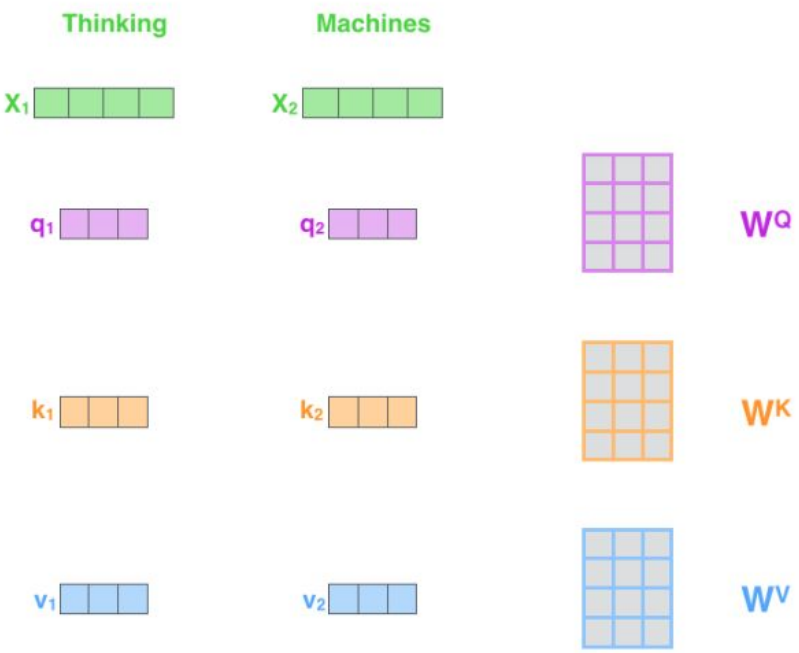
Attention

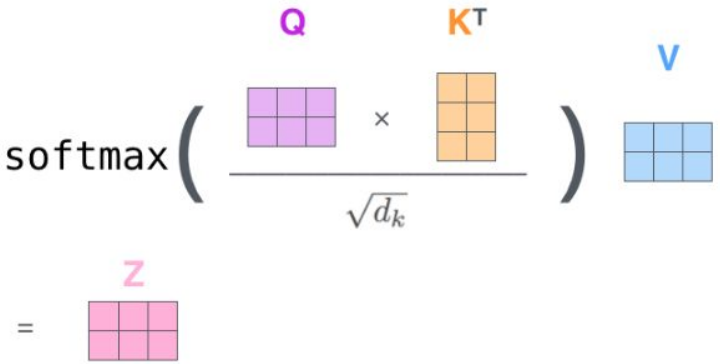

ENCODER #2
ENCODER #1
r1
r2
Feed Forward Neural Network
Feed Forward Neural Network
z1
z2
Self-Attention
x1
x2
Thinking
Machines

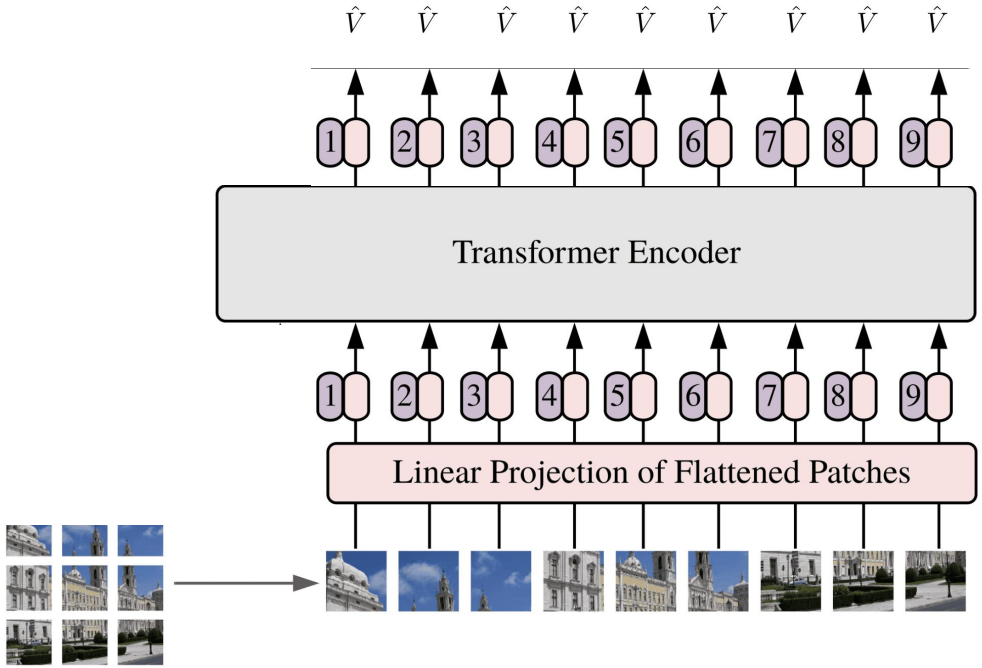
Attention

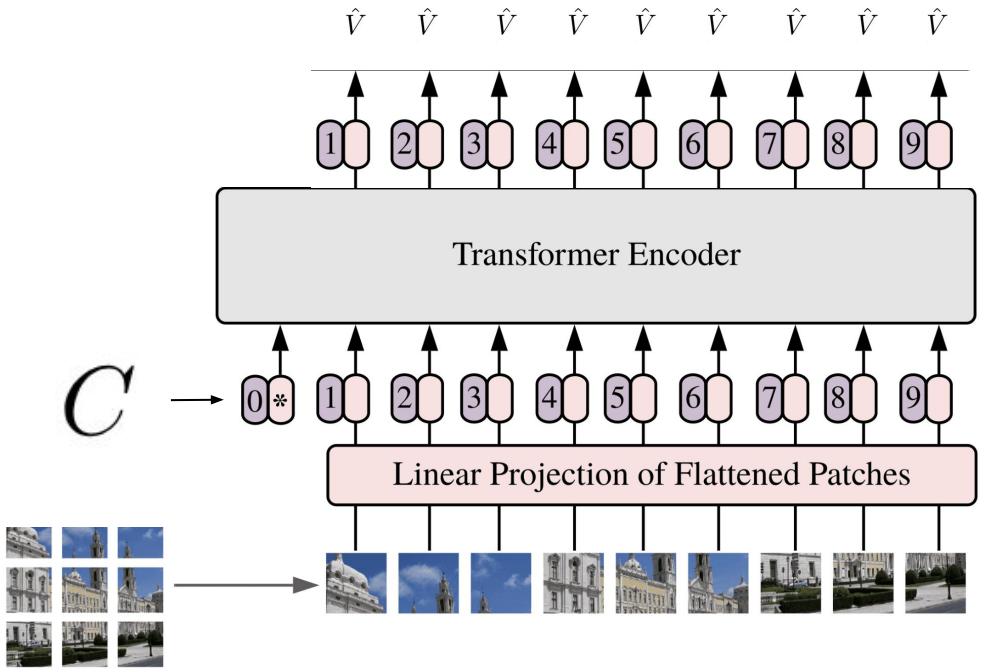


Attention


Thinking
Machines
X1
X2
q1
q2
k1
k2
v1
v2
WQ
WK
WV


Q
KT
V
softmax
×
√dk
Z
=


$\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$
1 2 3 4 5 6 7 8 9
Transformer Encoder
1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches


$\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$ $\hat{V}$
1 2 3 4 5 6 7 8 9
Transformer Encoder
C
0* 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches