Hyper-Transforming Latent Diffusion Models

Ignacio Peis
Technical University of Denmark
ipeaz@dtu.dk

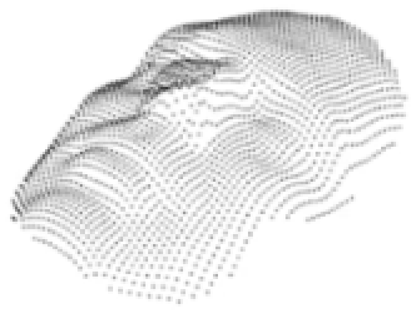
PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

Motivation

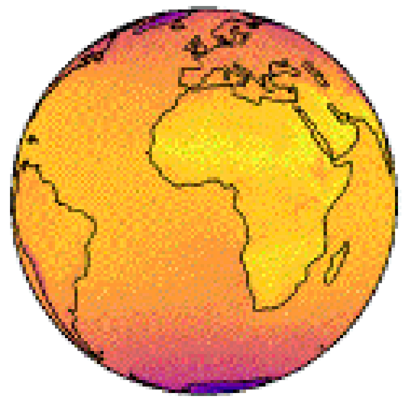
- We typically discretized data that are continuous in nature.

2D Images

3D Images

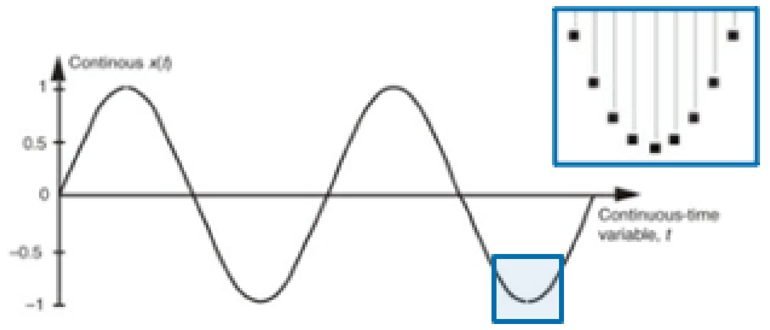


Polar data



Spatial

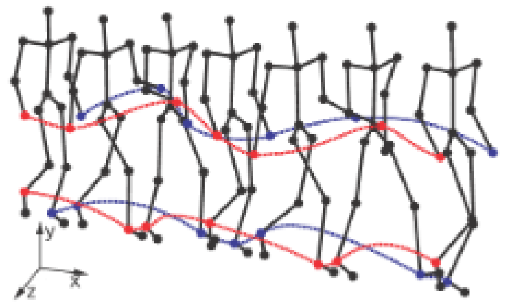
Time series



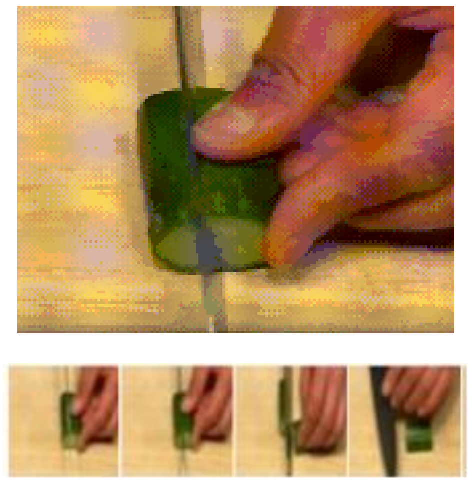
Audio

Temporal

Motion sequences



Video

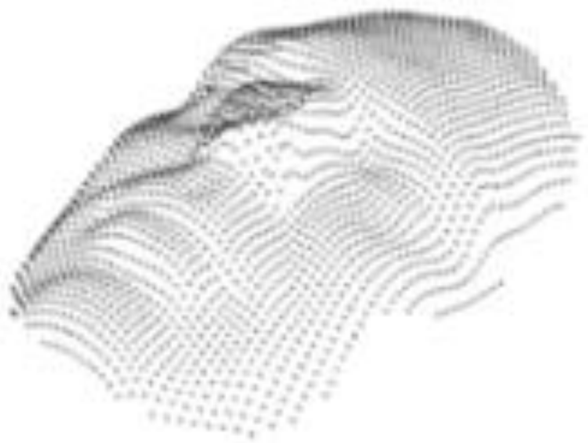


Spatio-temporal

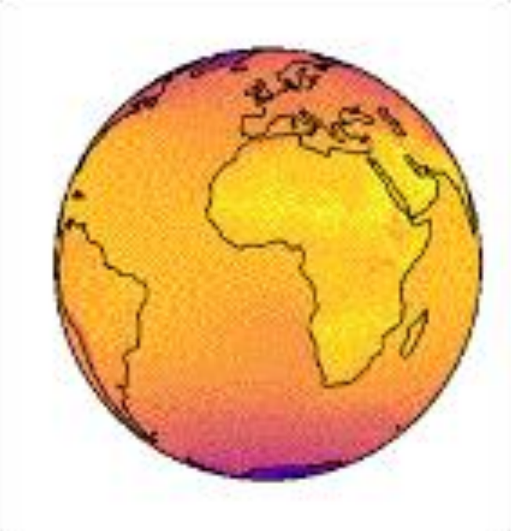
Motivation

- Real data can be expressed as a function over continuous coordinate systems.

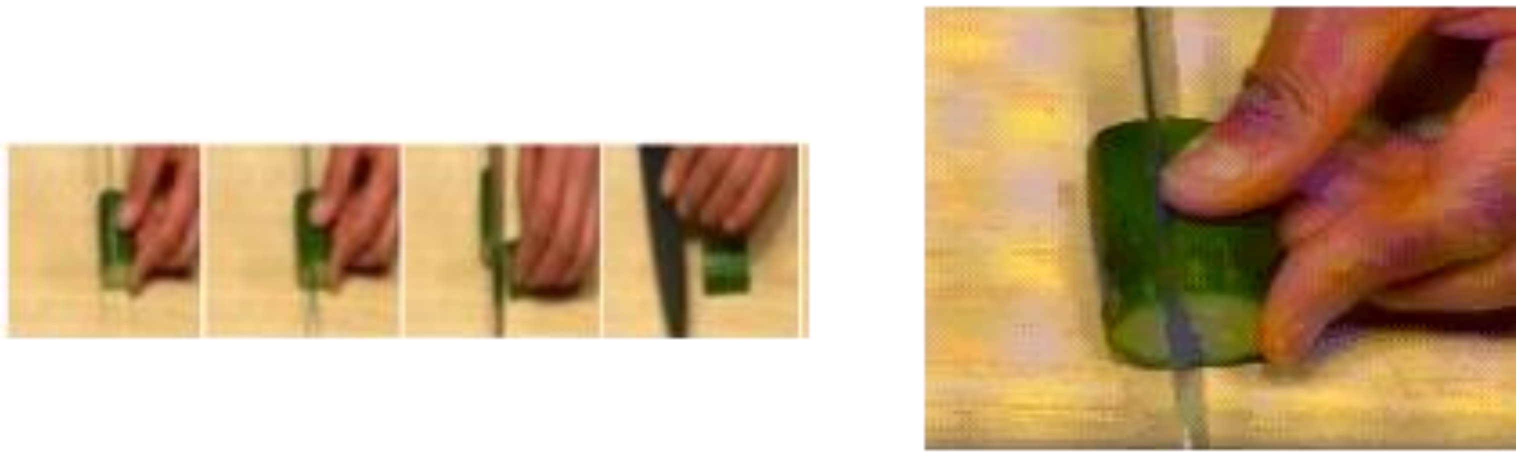
$f : \mathbb{R}^2 \to \mathbb{R}^3, f(x_1, x_2) = (r, g, b)$



$f : \mathbb{R}^3 \to \{0, 1\}, f(x_1, x_2, x_3) = p$



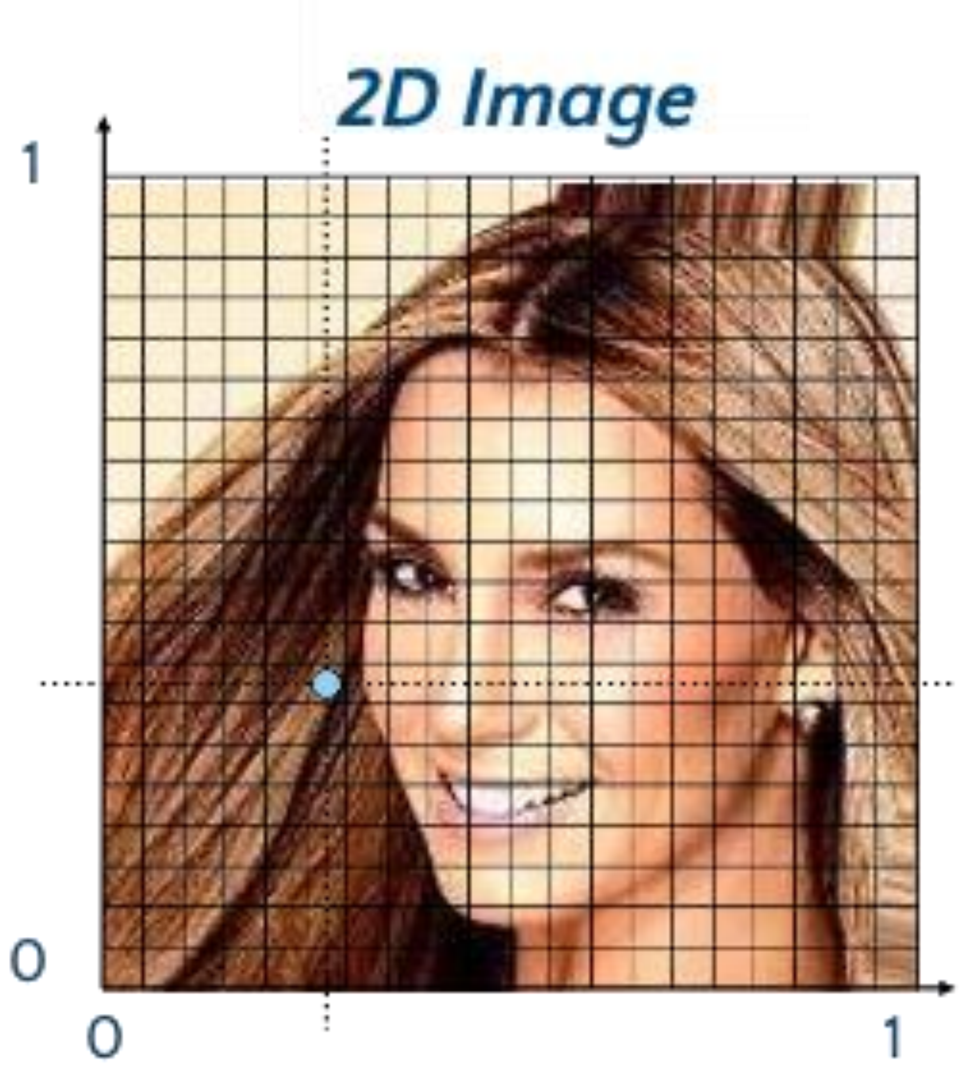
$f : \mathbb{R}^2 \to \mathbb{R}, f(\varphi, \lambda) = T$



$f : \mathbb{R}^3 \to \mathbb{R}^3, f(x_1, x_2, t) = (r, g, b)$

Motivation

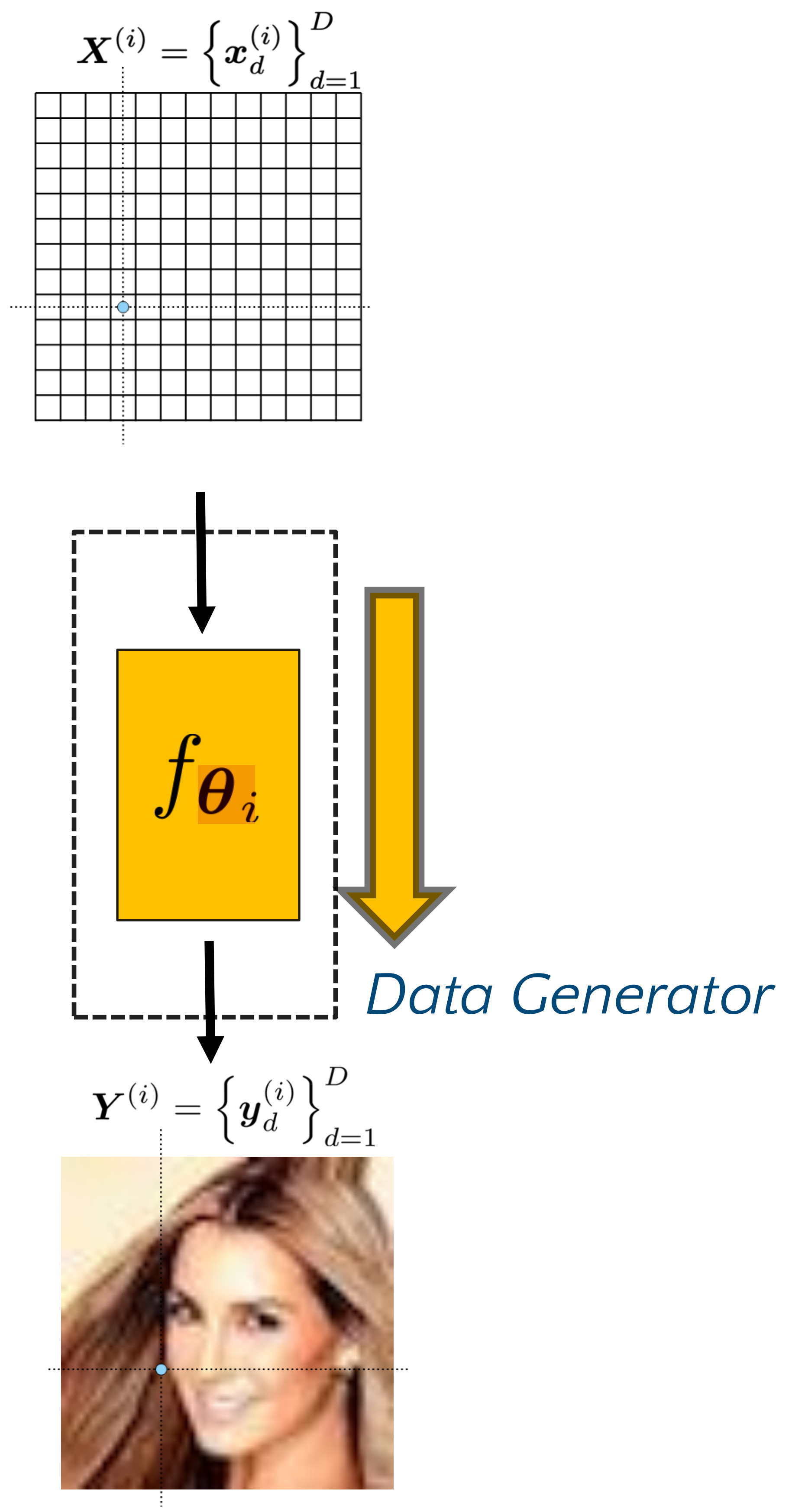
- Focusing on images:



- Generator function $f : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ creates this speficic image with the mapping $f(\boldsymbol{x}_d) = \boldsymbol{y}_d$, $d \in [1, ..., D]$
- Each pixel is now a pair $\{\boldsymbol{x}_d, \boldsymbol{y}_d\}$ where $x_d \in \mathbb{R}^2$, $y_d \in \mathbb{R}^3$
- Full image is a pair of sets $\boldsymbol{X} = \{\boldsymbol{x}_d\}_{d=1}^{D}$, $\mathbf{Y}_d = \{\boldsymbol{y}_d\}_{d=1}^{D}$

Implicit Neural Representations

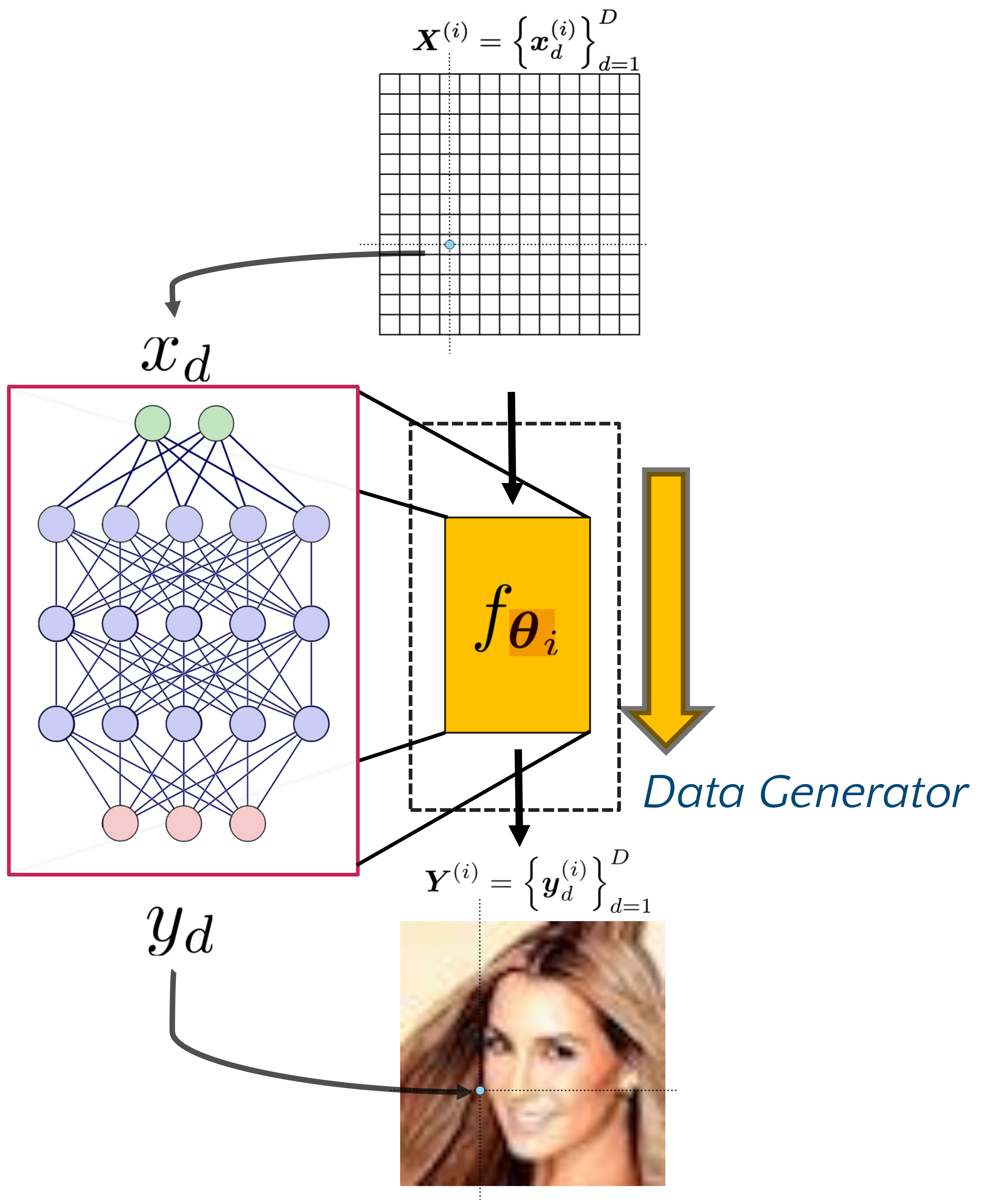
INRs [20-22]



[20] Sitzmann et at., 2020 [21] Mescheder et at., 2019 [20] Sitzmann et at., 2019

Implicit Neural Representations

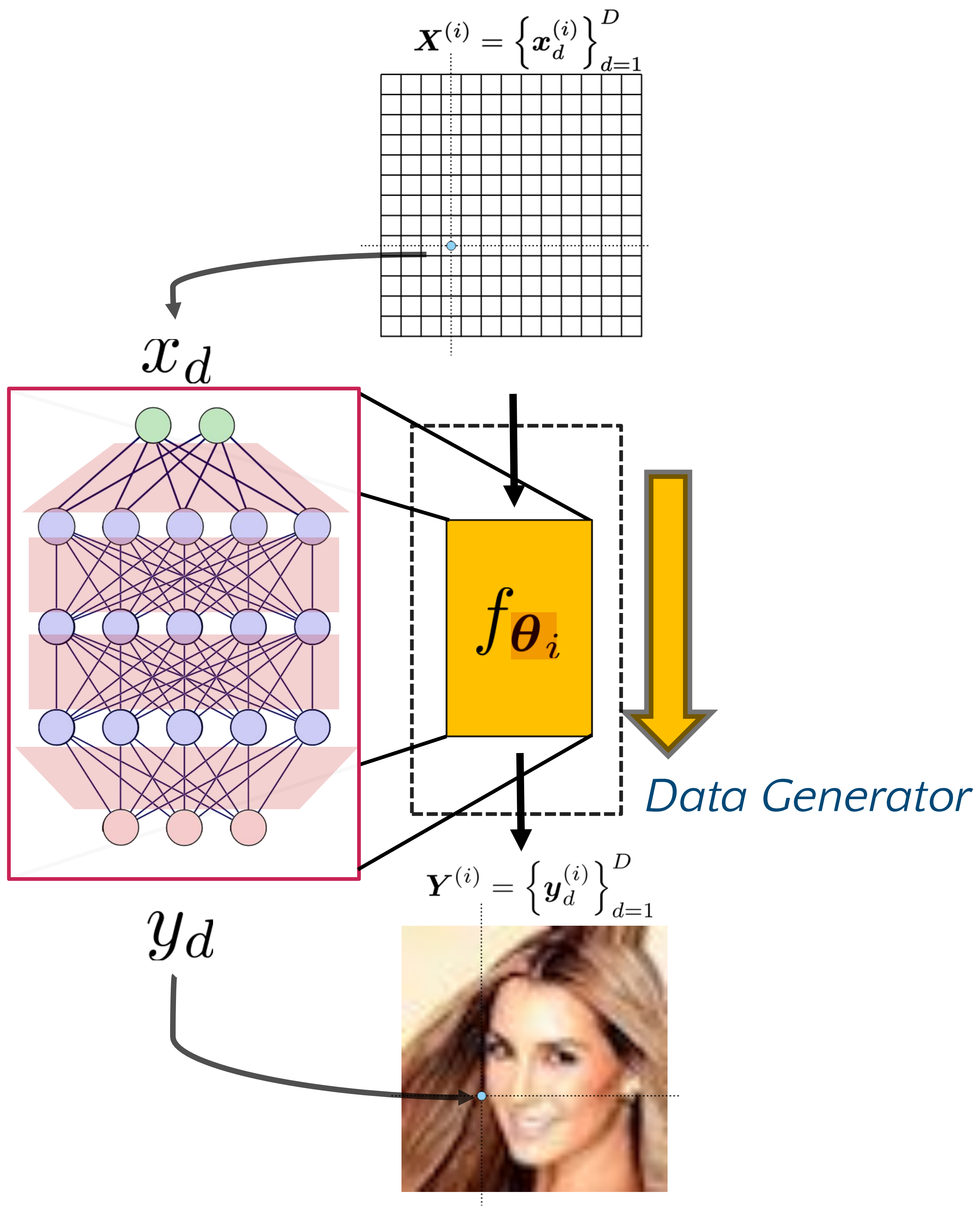
INRs [20-22]



[20] Sitzmann et at., 2020 [21] Mescheder et at., 2019 [20] Sitzmann et at., 2019

Implicit Neural Representations

INRs [20-22]

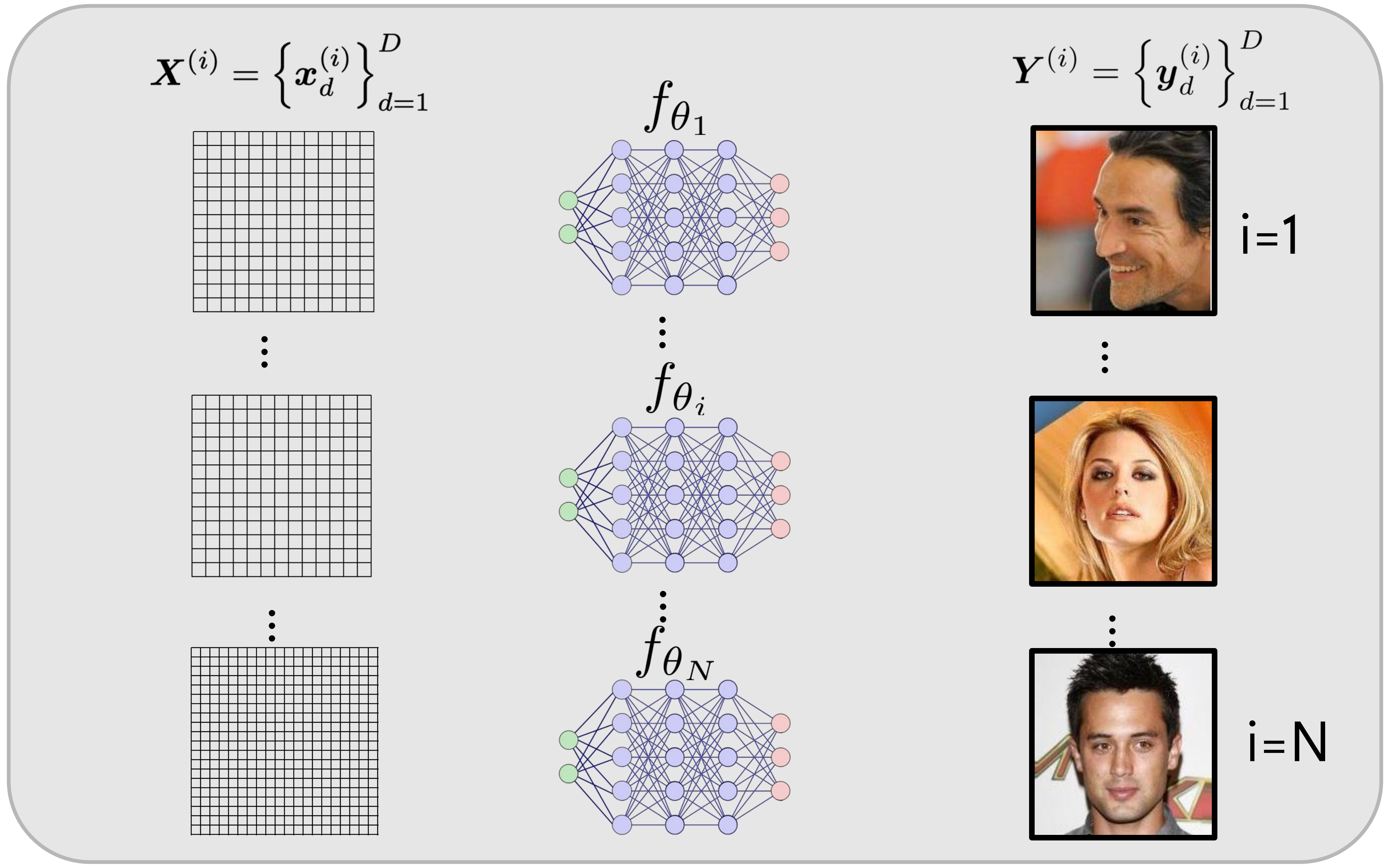


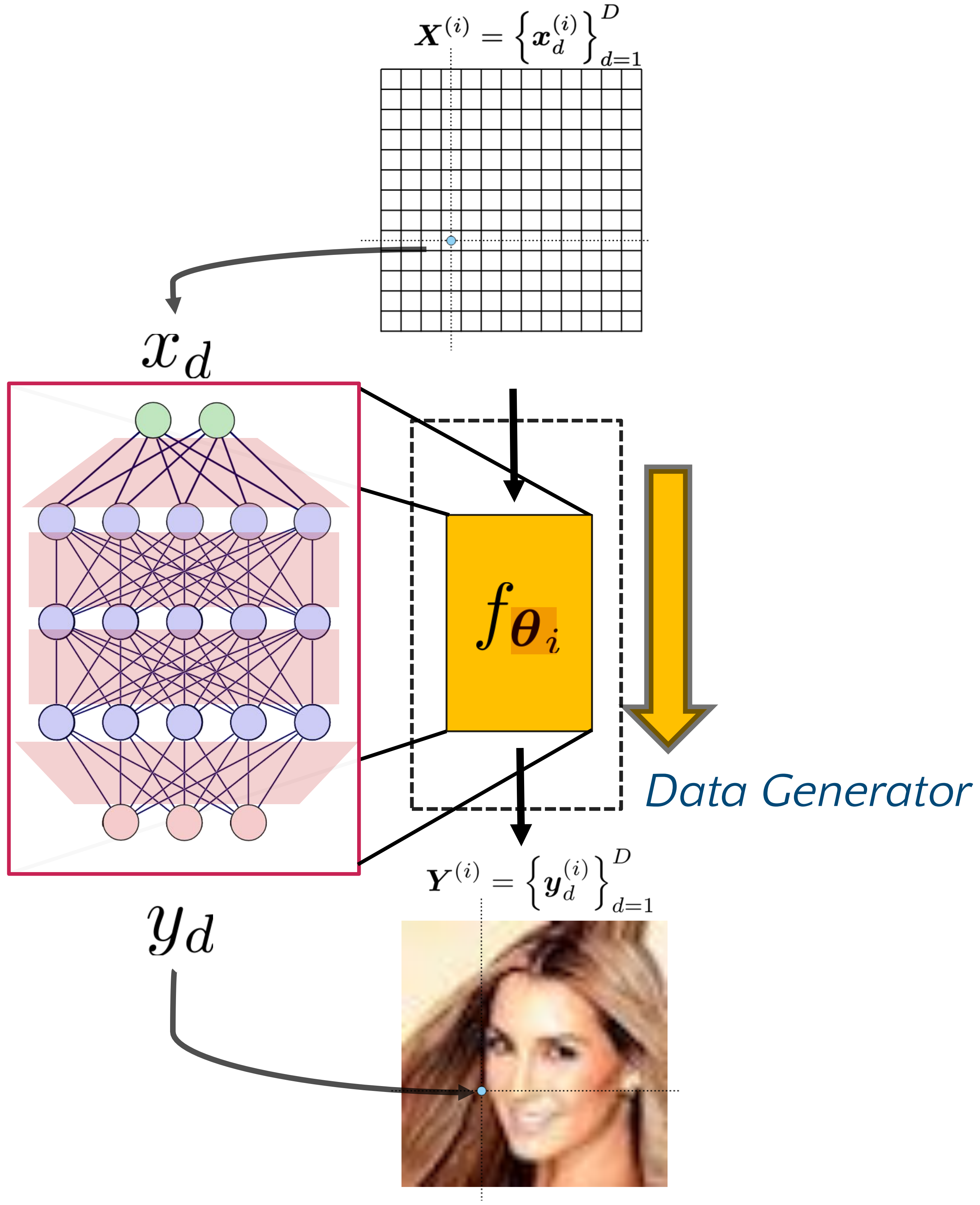
[20] Sitzmann et at., 2020 [21] Mescheder et at., 2019 [20] Sitzmann et at., 2019

Implicit Neural Representations

INRs [20-22]

Data generator $f_{\boldsymbol{\theta}_i}$ is unique to each image





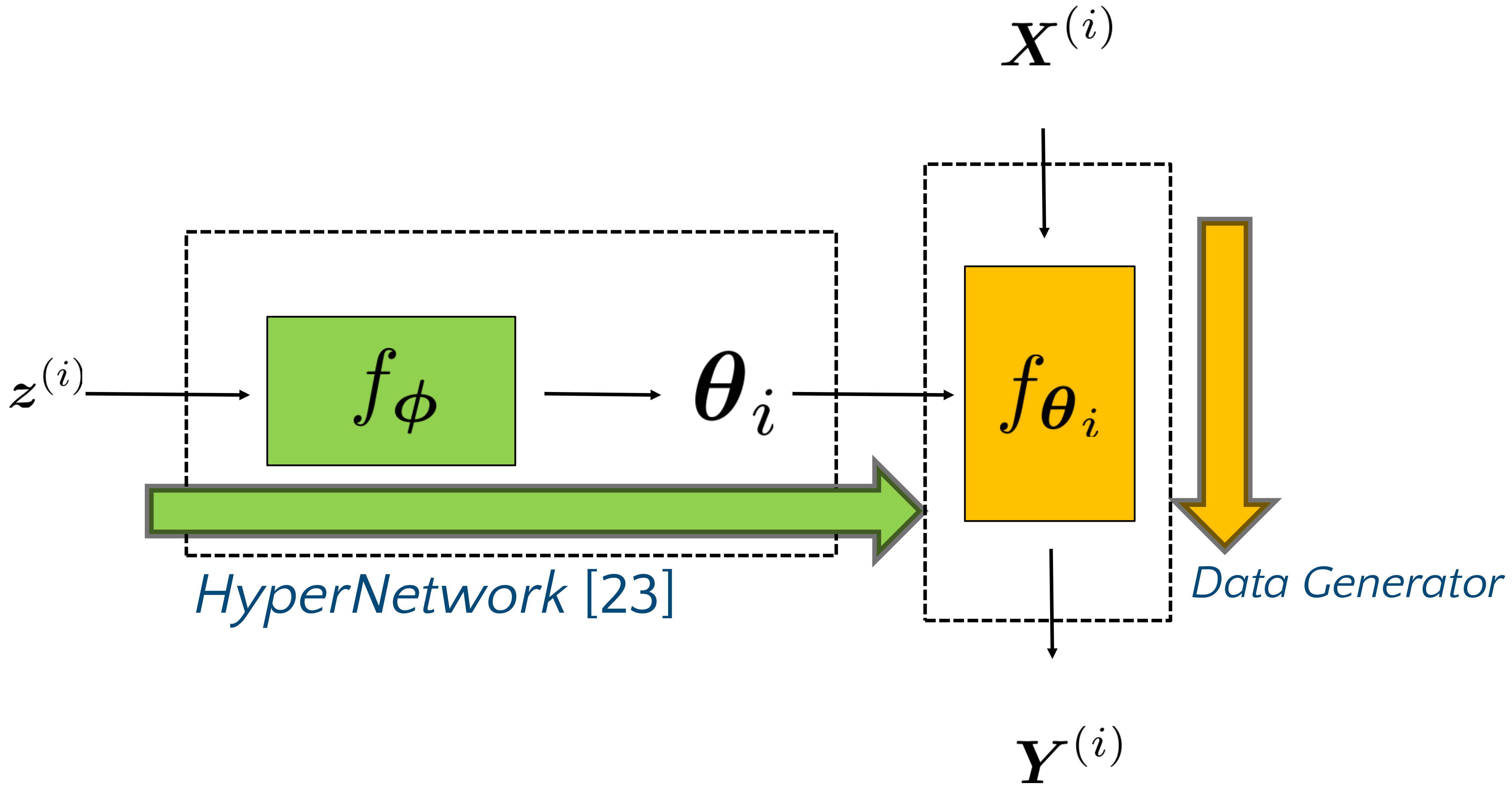
[20] Sitzmann et at., 2020 [21] Mescheder et at., 2019 [20] Sitzmann et at., 2019

Deep Generative Models of INRs

How to scale to large datasets?

How to map a latent representation to an INR?

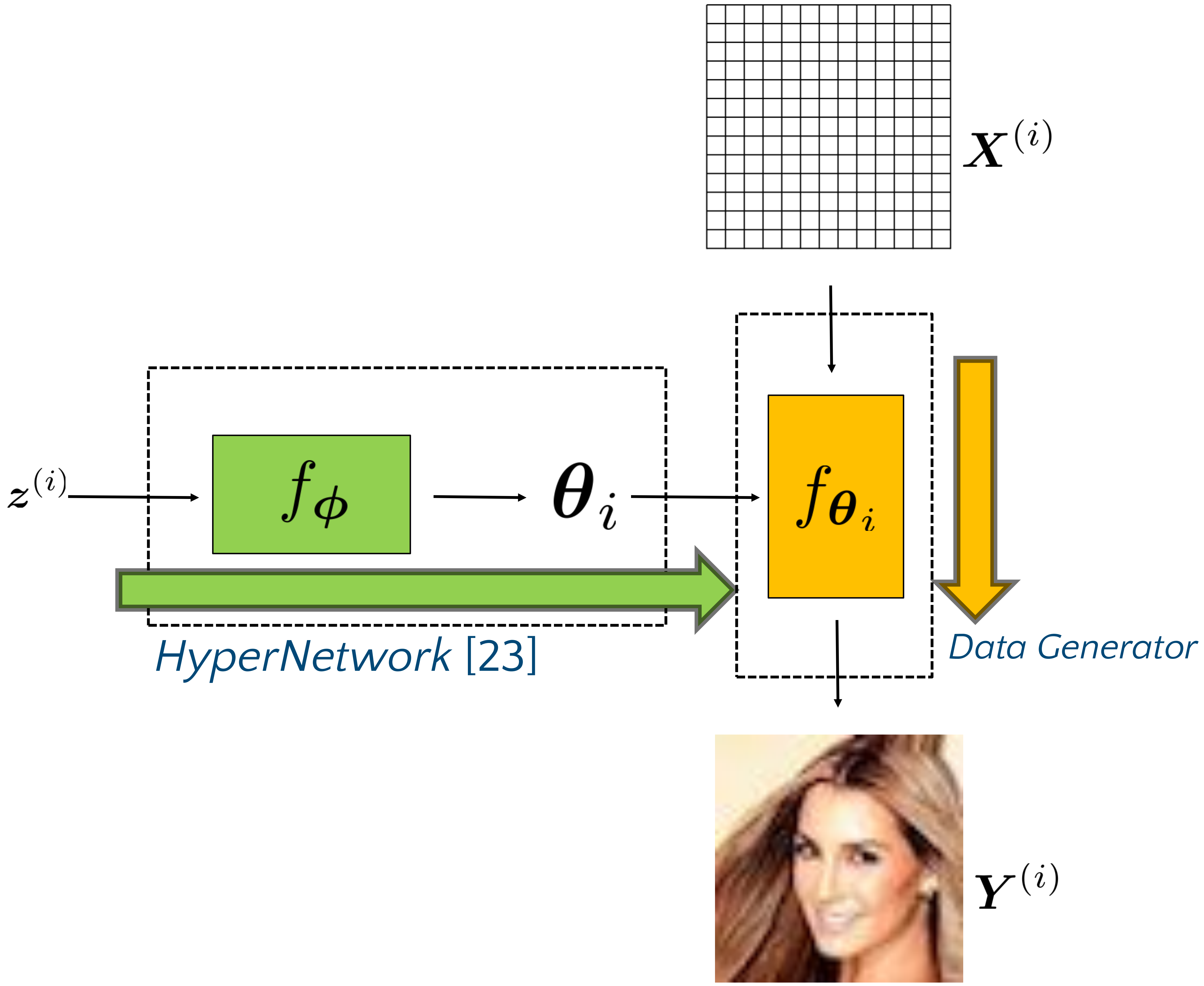
Deep Generative Models of INRs



[23] Ha et at., 2017

Deep Generative Models of INRs

Have $\boldsymbol{z}^{(i)}$, a summary representation of image.



[23] Ha et at., 2017

Previous work

GASP[5]

- Adversarial training:

Point-wise Convolution

Figure 5: Training procedure for GASP: 1. Sample a function and evaluate it at a set of coordinate locations to generate fake point cloud. 2. Convert real data sample to point cloud. 3. Discriminate between real and fake point clouds.

✗ Can’t tackle inference related tasks.

[5] Dupont et at., 2020

Previous work

Functa[6]

- Decoupled training:

    1. Fit an INR per datapoint using SIREN[20] and **modulation vectors**, named **functas**.

    2. Train any generative model on the functa dataset of vectors.

✗ Computationally expensive inference.

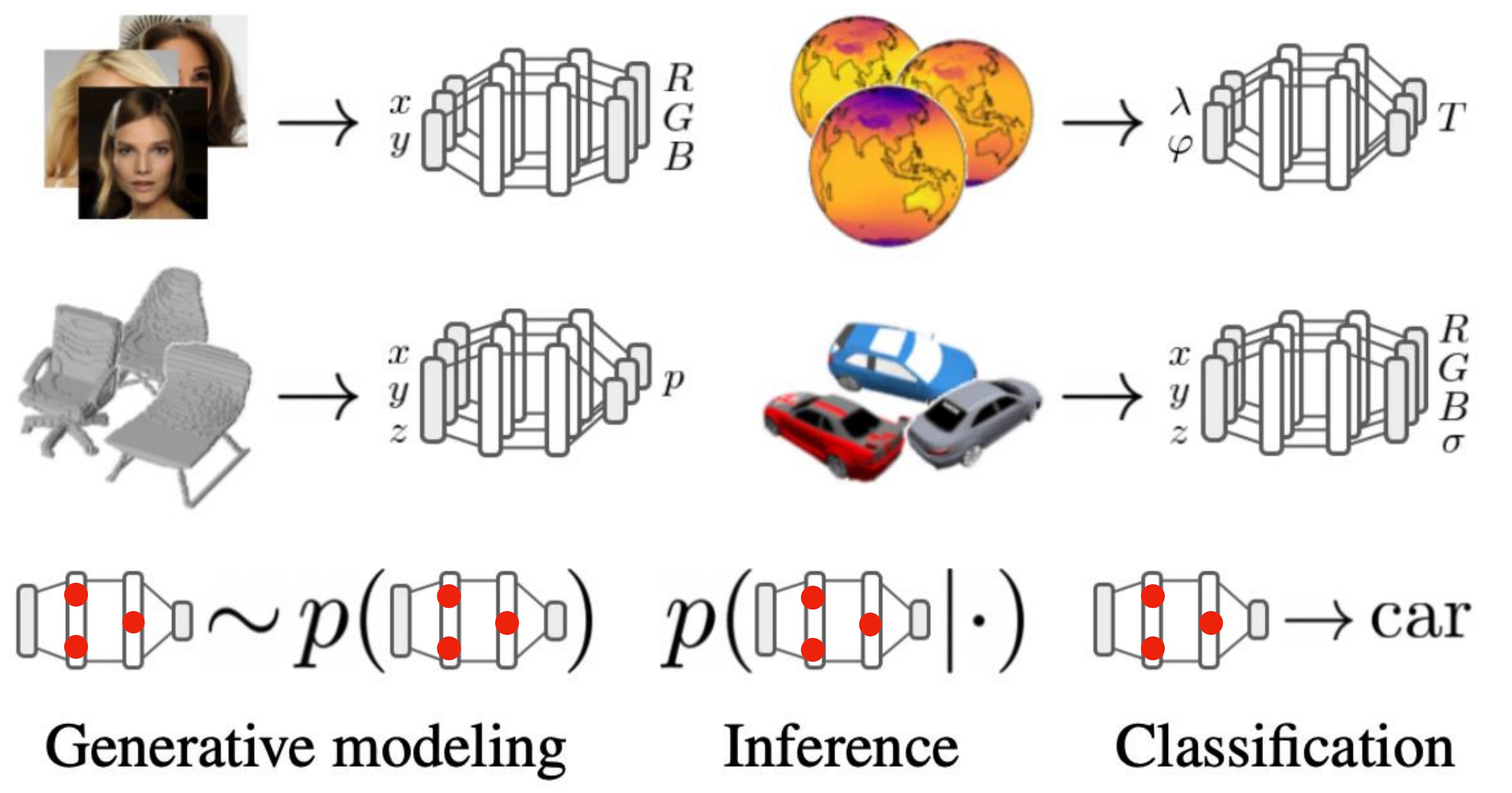


Figure 1. We convert array data into functional data parameterized by neural networks, termed *functa*, and treat these as data points for various downstream machine learning tasks.

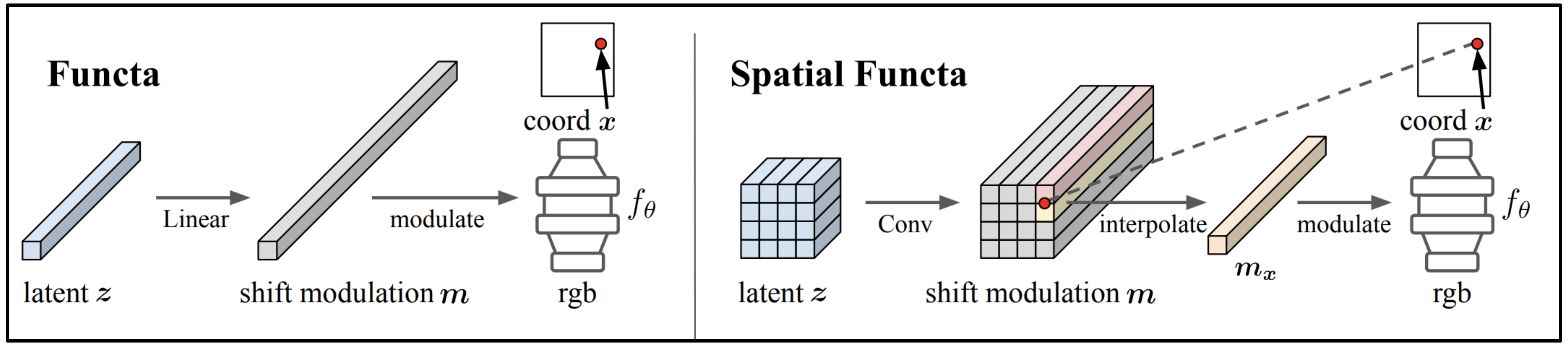
[6] Dupont et at., 2022 [20] Sitzmann et at., 2020

Previous work

Spatial Functa[26]

- Decoupled training:
  1. Fit an INR per datapoint using SIREN[20] and **modulation tensor**.
  2. Train any generative model on the functa dataset of tensors.

✗ Computationally expensive inference.



[26] Bauer et at., 2023 [20] Sitzmann et at., 2020

Deep Generative Models of INRs

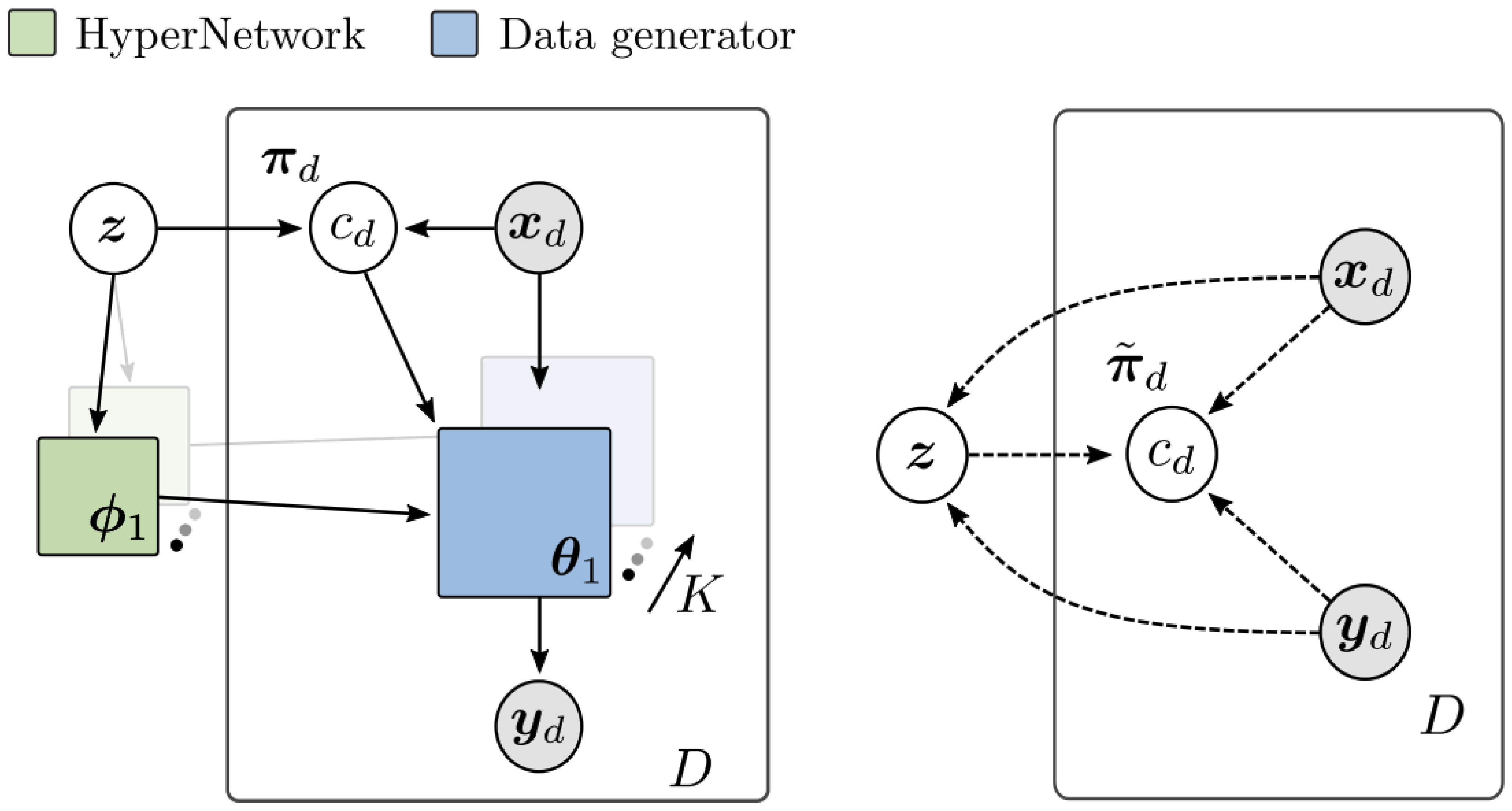
How to infer the latent representation ***z***?

$$q_\gamma(\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{X}) \qquad p_\psi(\boldsymbol{z})$$

Proposed methods (1)

VAMoH

Variational Mixture of HyperGenerators [25]

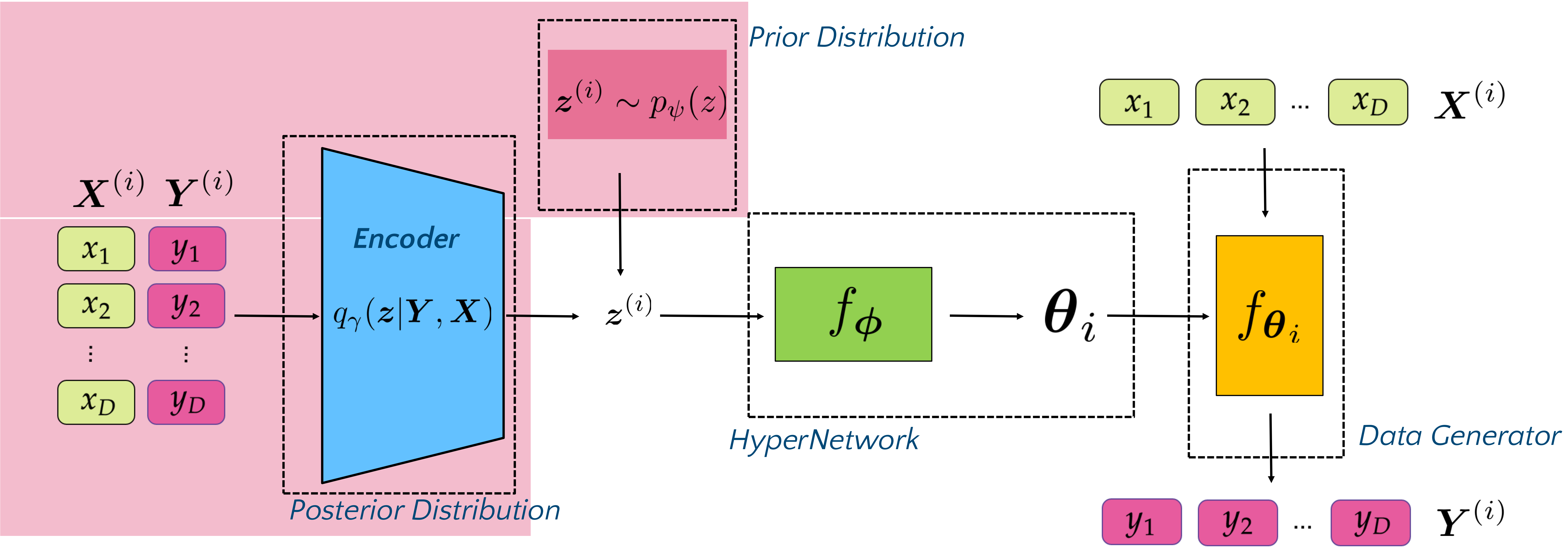


(a) Generative model (b) Inference model

[25] Koyuncu et at., 2023

VAMoH

Encoder

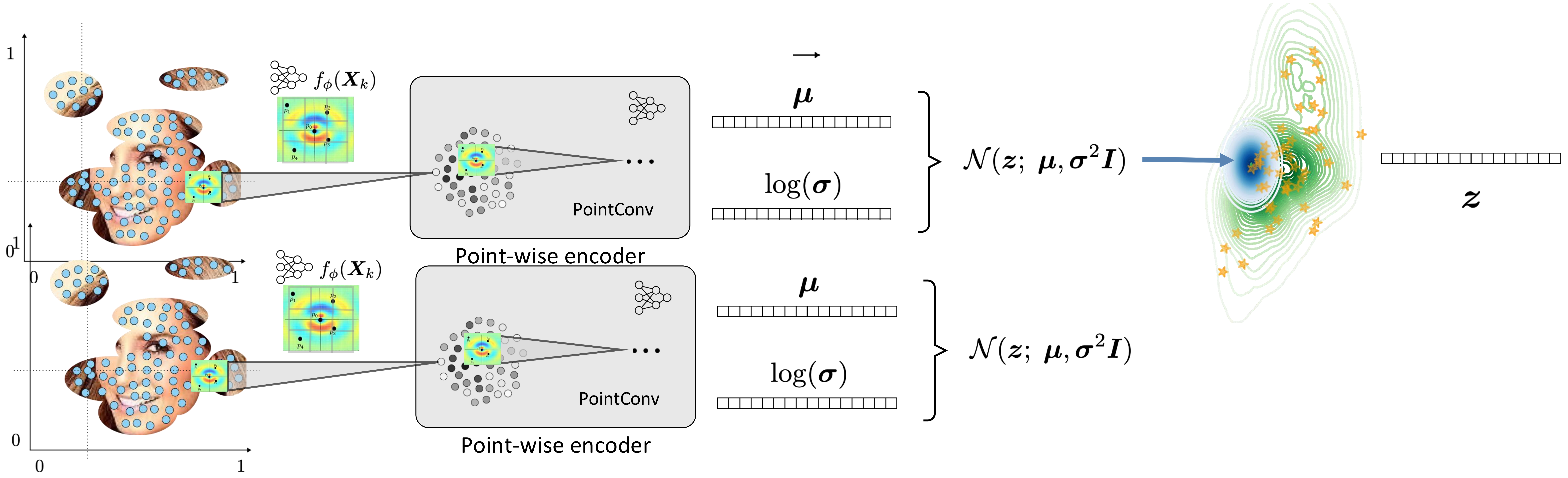


$z^{(i)}$: Latent Variable

VAMoH

Encoder

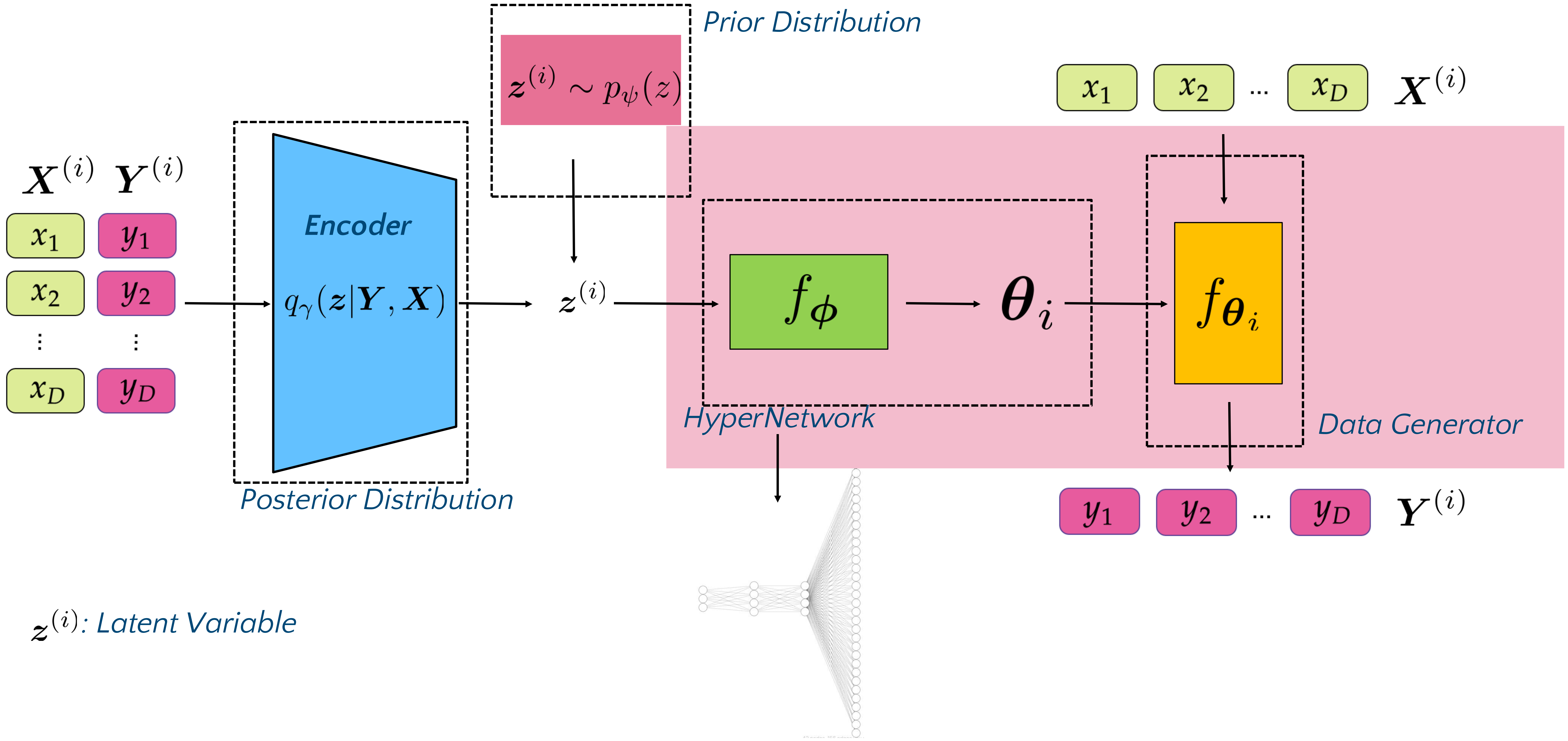
- PointConv[21] encoder for point clouds.



[21] Wu et at., 2019

VAMoH

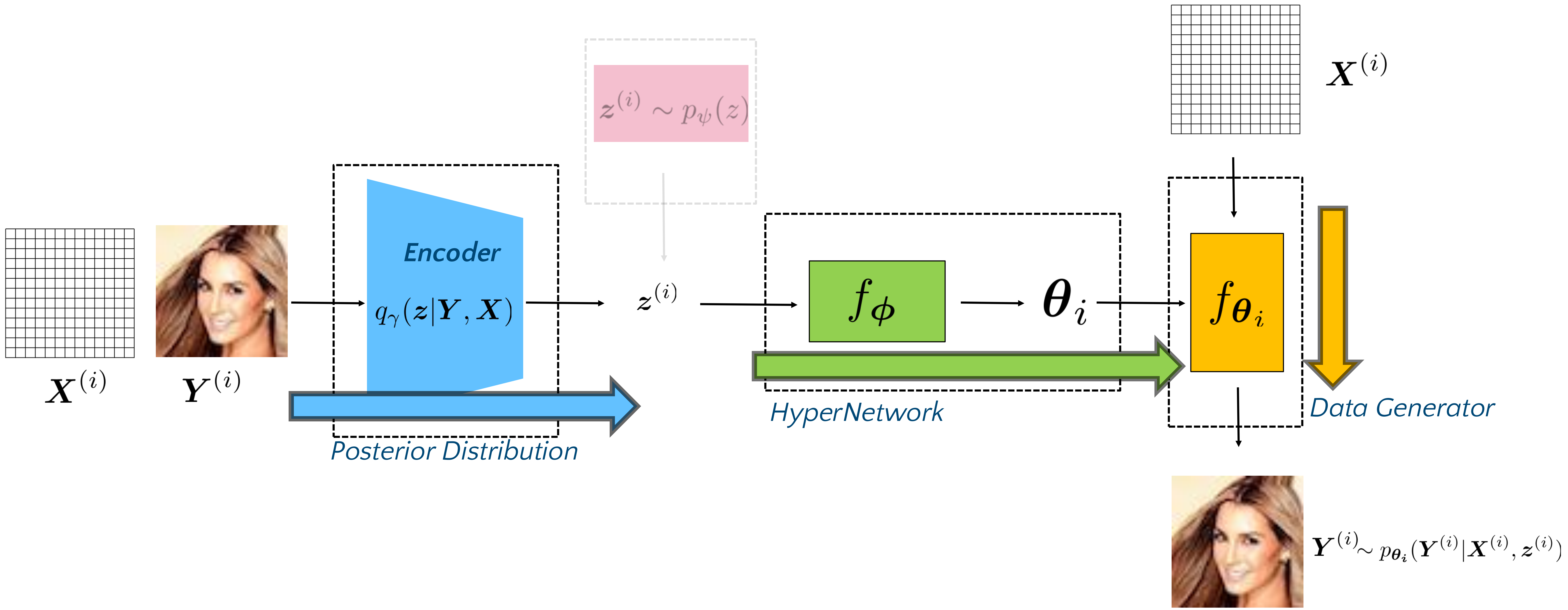
Decoder



$z^{(i)}$: *Latent Variable*

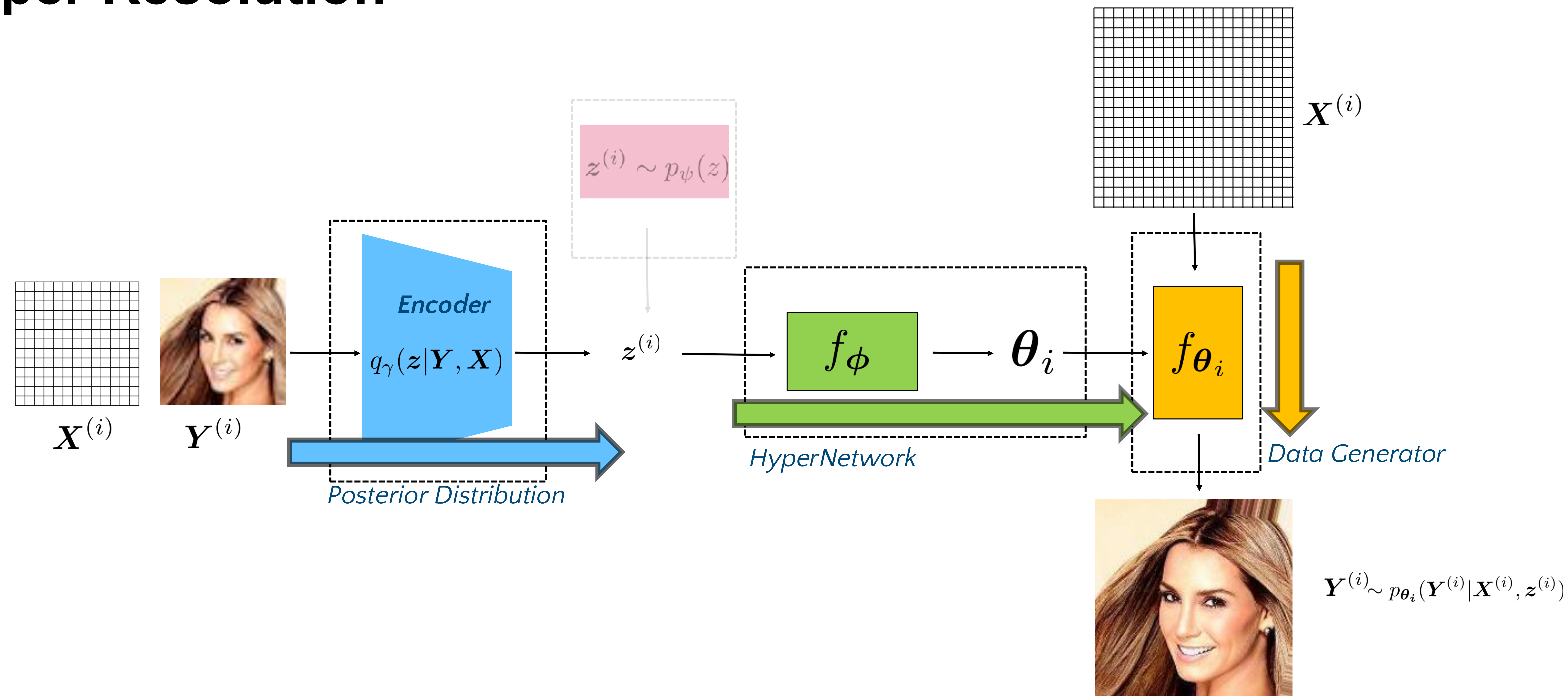
VAMoH

Reconstruction



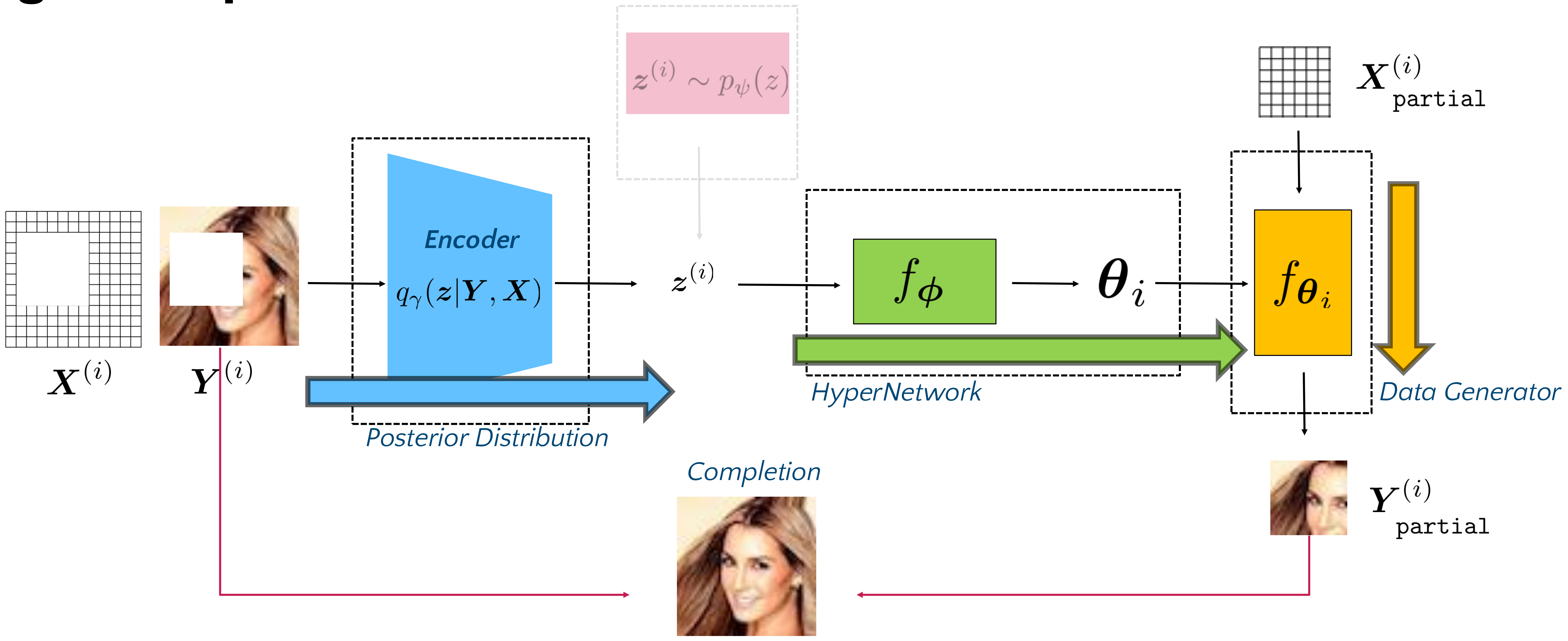
VAMoH

Super Resolution



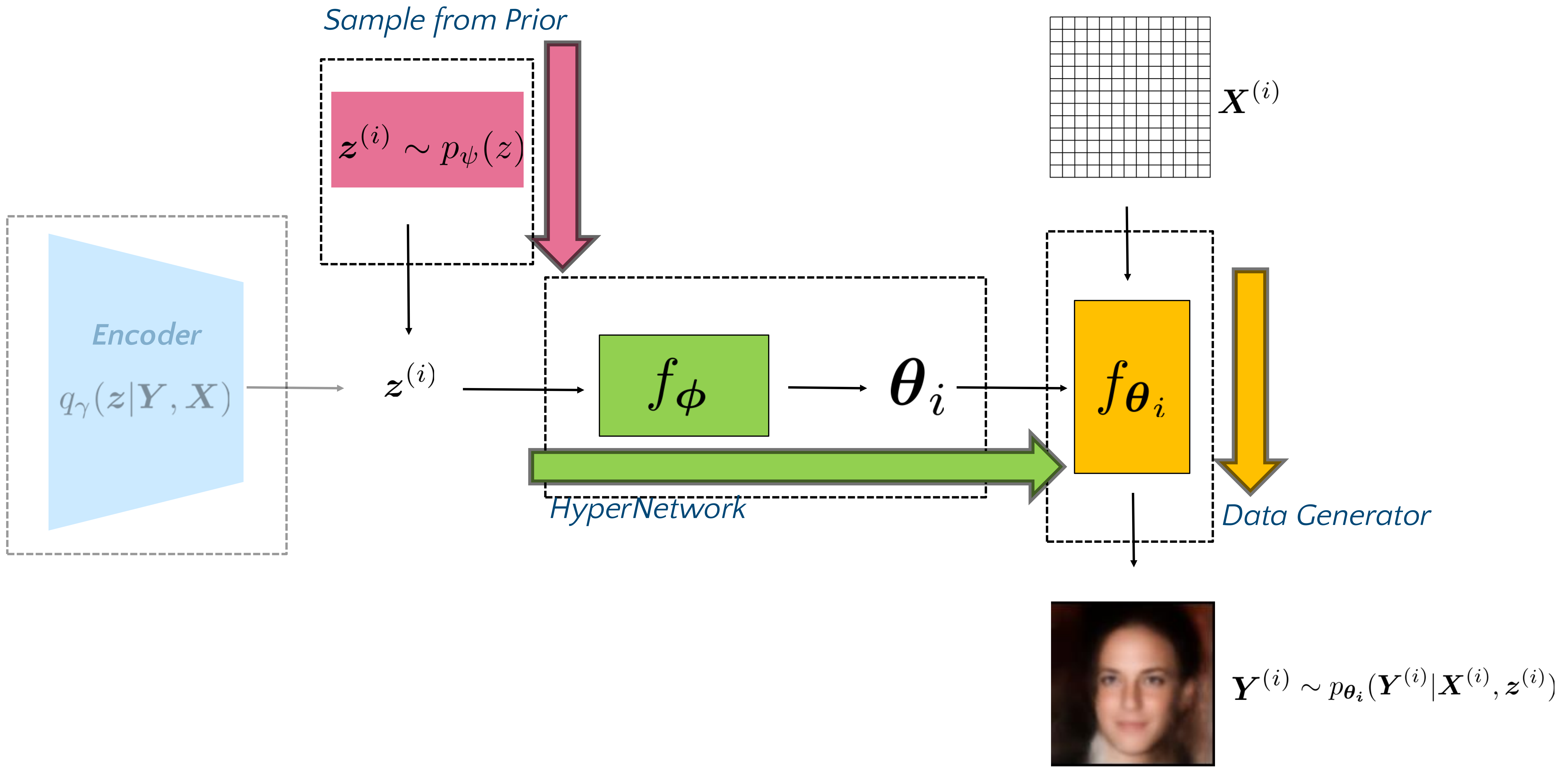
VAMoH

Image Completion



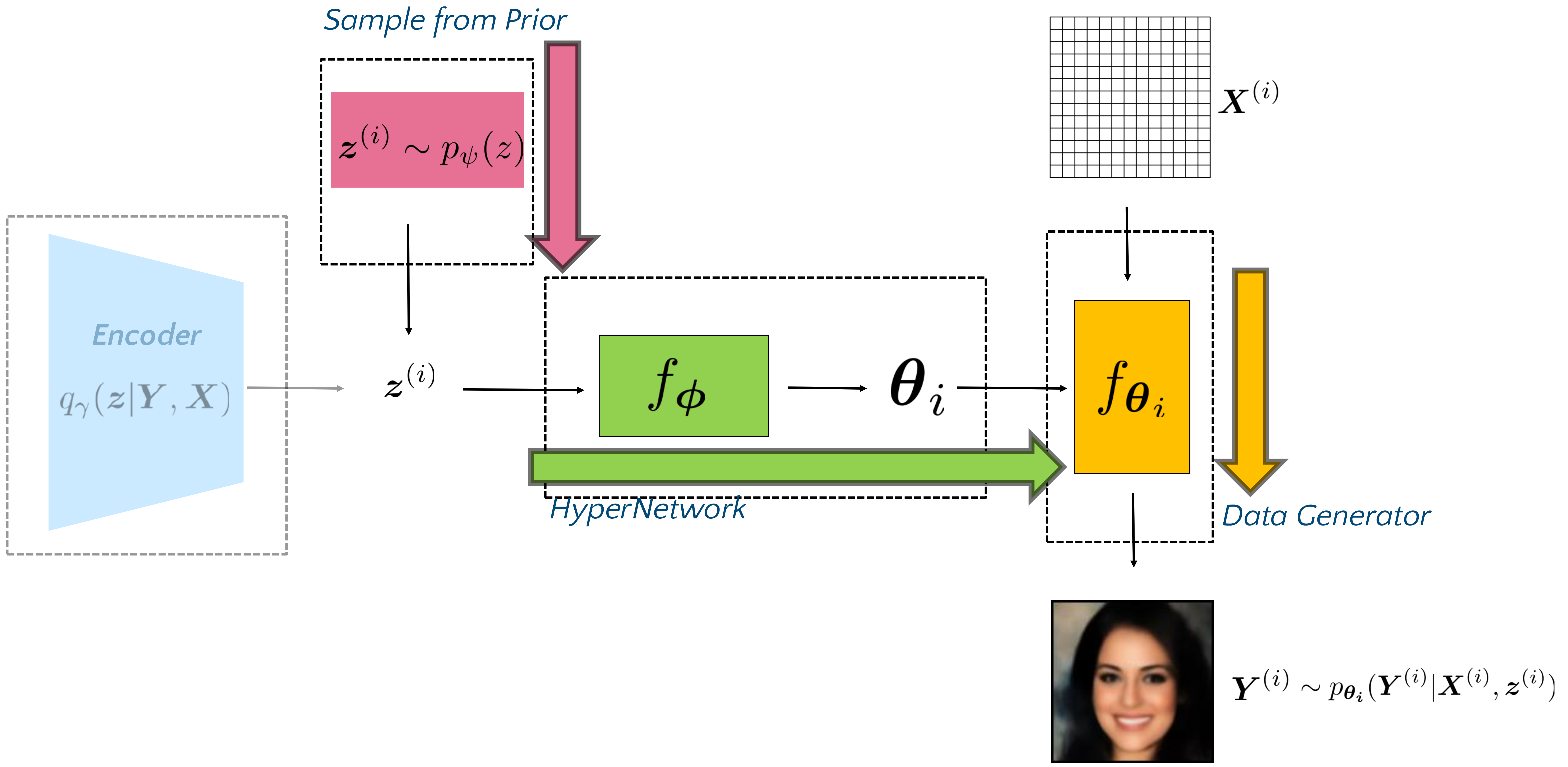
VAMoH

Image Generation



VAMoH

Image Generation



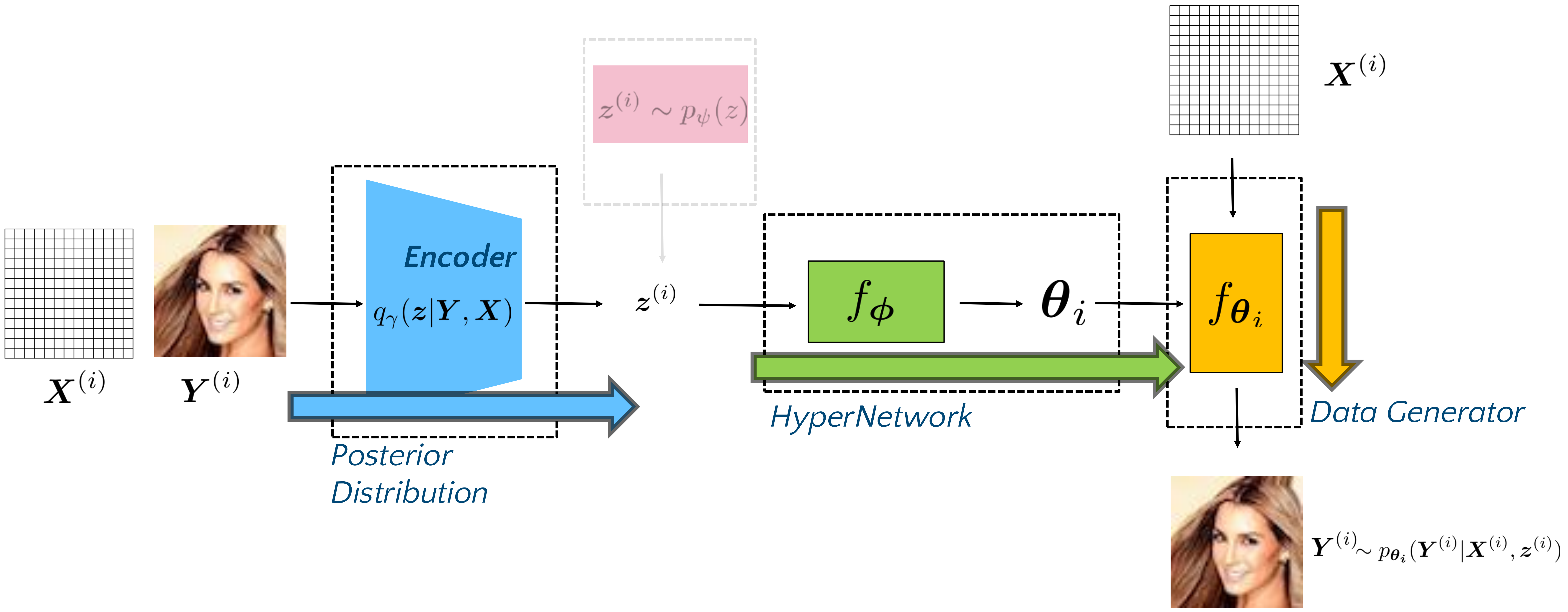
VAMoH

Optimization

How to learn all these steps end-to-end from data?

VAMoH

Optimization



Generative model:

$$p(Y, z \mid X) = p_\theta(Y \mid X, z)p_\psi(z)$$

Aim: Learning latent variables z given data (intractable) objective)

$$p(z|Y, X)$$

How: Learn an approximation

$$q_\gamma(z|Y, X) \approx p(z|Y, X)$$

VAMoH

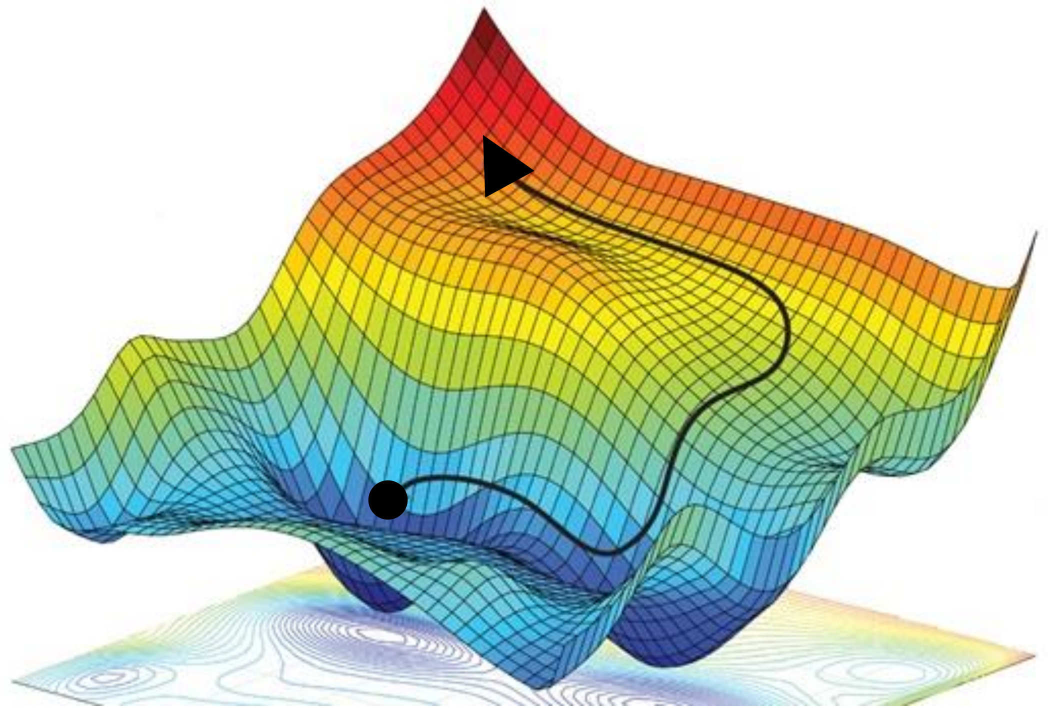
Optimization

- For a single data sample $(\boldsymbol{X}, \boldsymbol{Y})$

$$\max_{\phi,\gamma} \mathcal{L}(\phi,\gamma;\boldsymbol{Y},\boldsymbol{X}) = \underbrace{\boxed{\max_{\phi,\gamma} \mathbb{E}_{q_{\boldsymbol{\gamma}}(\boldsymbol{z}|\boldsymbol{Y},\boldsymbol{X})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y} \mid \boldsymbol{X},\boldsymbol{z})\right]}}_{\text{Reconstruction}} - \underbrace{\boxed{D_{KL}(q_\gamma(\boldsymbol{z}|\boldsymbol{Y},\boldsymbol{X}) \,\|p_\psi(\boldsymbol{z}))}}_{\text{Regularization}}$$

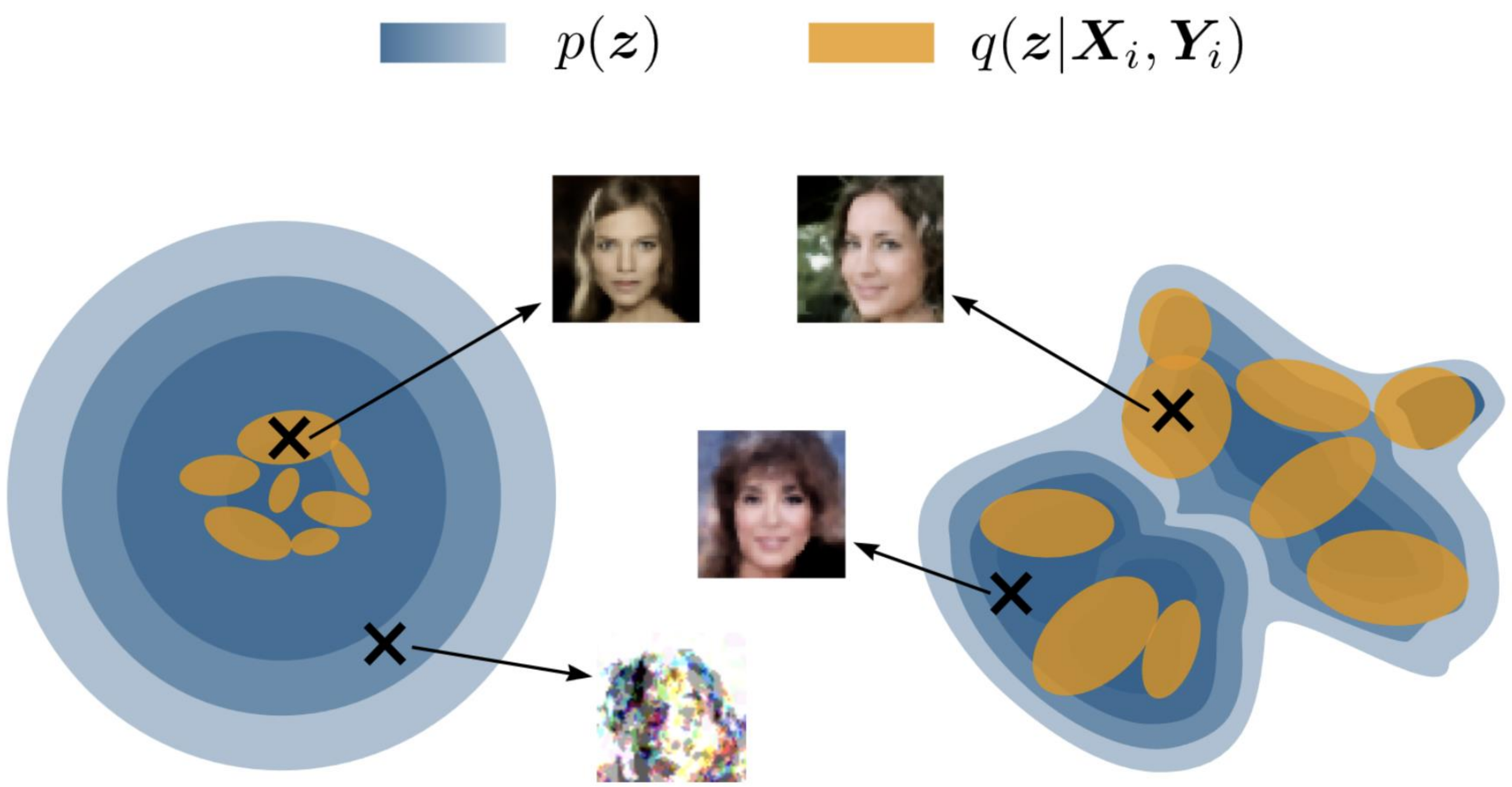
- For all samples in our dataset $\left(\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}\right), i \in [N]$

$$\max_{\phi,\gamma} \sum_{i=1}^{N} \mathcal{L}(\phi,\gamma;\boldsymbol{Y}^{(i)},\boldsymbol{X}^{(i)})$$



VAMoH

‘*Holes*’ problem



Regularization Term:
$\min_{\gamma} D_{KL}\left(q_\gamma(\boldsymbol{z} \mid \boldsymbol{Y}, \boldsymbol{X}) \| p_\psi(\boldsymbol{z})\right)$

We need to align the approximate posterior with the prior.
$p_\psi(\boldsymbol{z}) \quad q_\gamma(z)$

Problem:

If the prior is too simple, it hinders generation quality.

Solution:

Learn a more complex $p_\psi(z)$ with another NN.

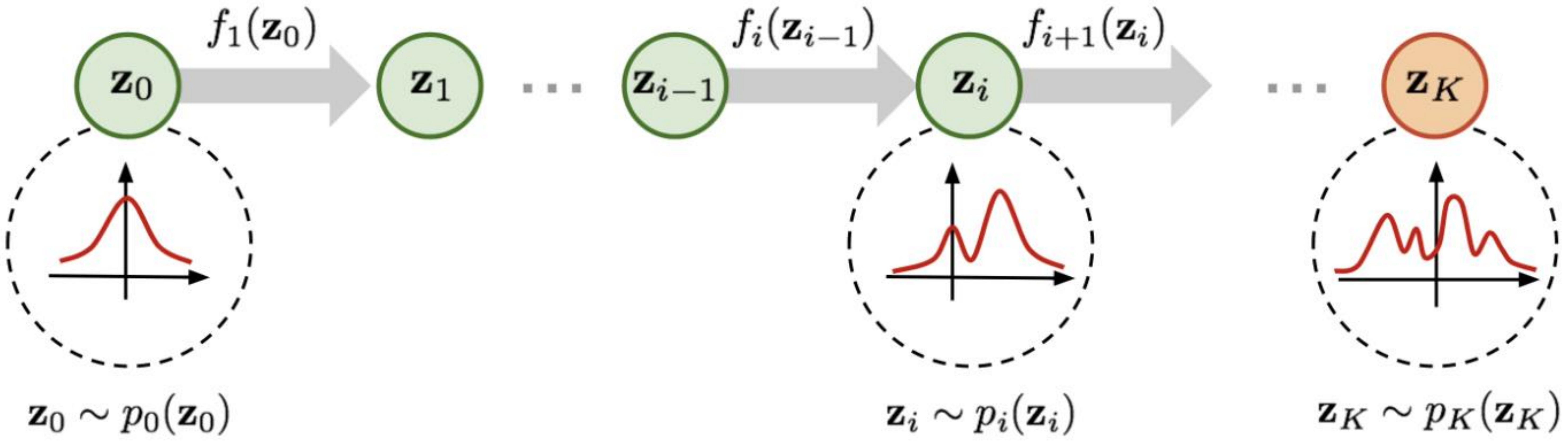
$$\min_{\gamma,\psi} D_{KL}(q_\gamma(\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{X}) \parallel p_\psi(\boldsymbol{z}))$$

VAMoH

Flow-based prior

- More expressive prior using RealNVP (Real-valued, Non-Volume Preserving) Flow.

[23]

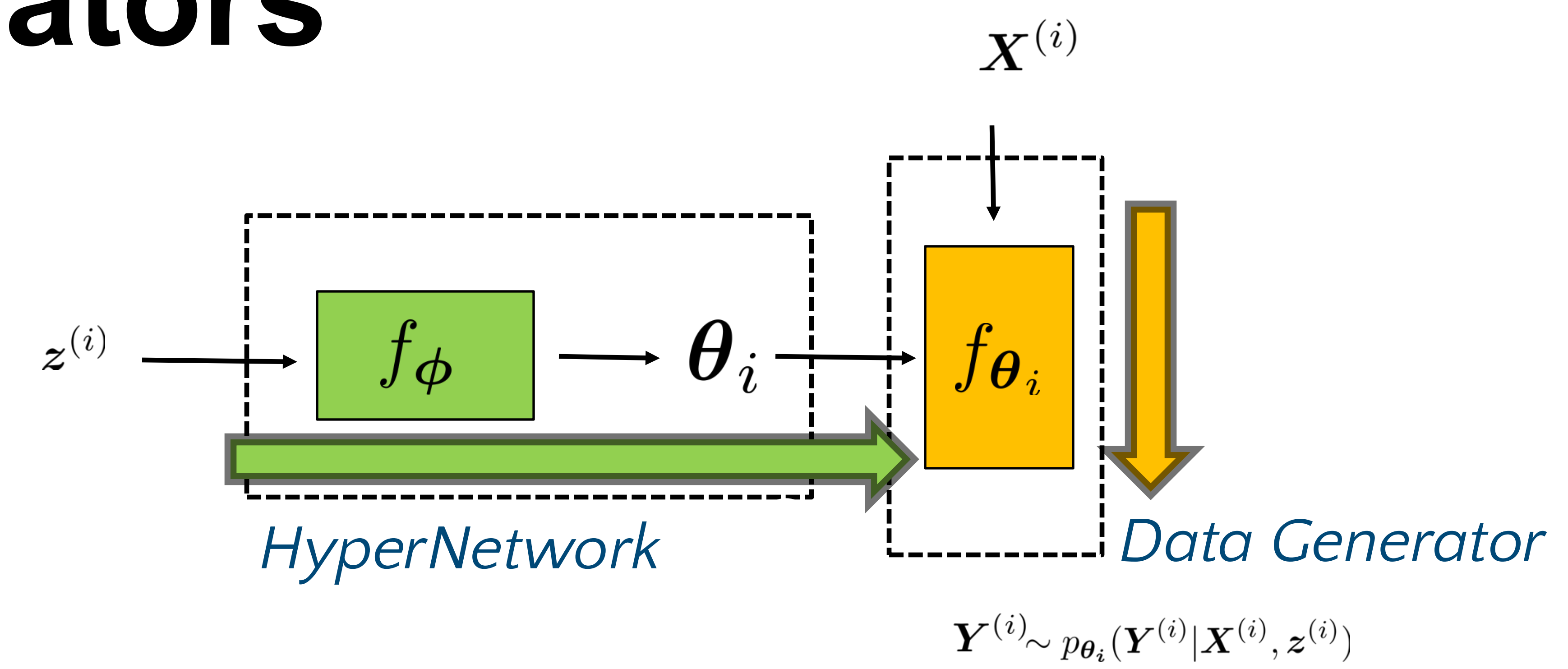


$$\boldsymbol{z}^{(i)} \sim p_\psi(z)$$

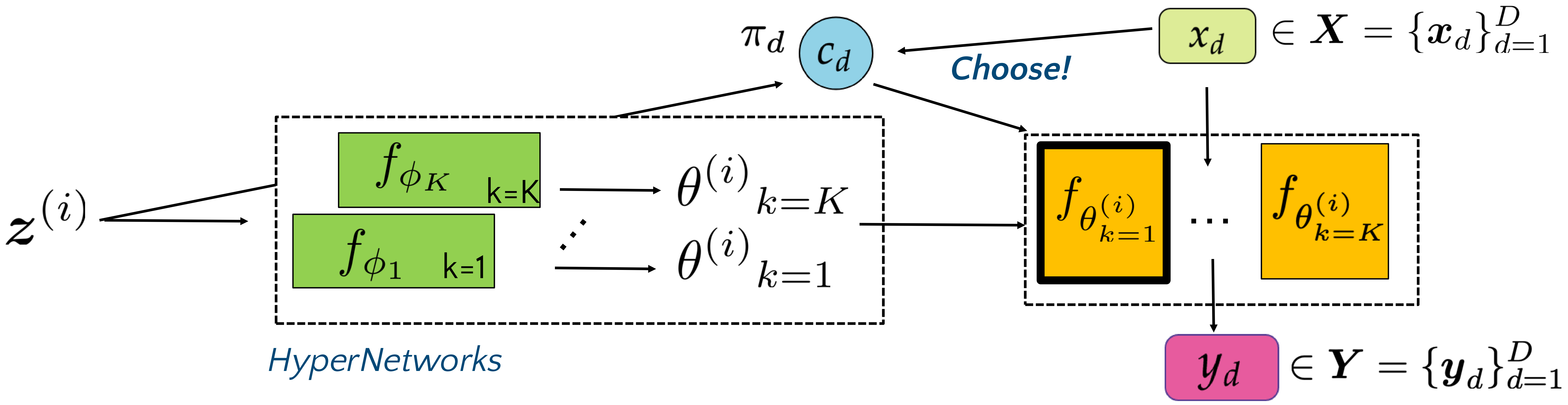
VAMoH

Mixture of HyperGenerators

Single HyperGenerator



Mixture of HyperGenerators



VAMoH

Mixture of HyperGenerators

Image Reconstruction with Mixture of HyperGenerators

VAMoH

- For a single data sample $(\boldsymbol{X}, \boldsymbol{Y})$

$$\mathcal{L}(\boldsymbol{Y}, \boldsymbol{X}; \psi, \phi, \boldsymbol{\gamma}) = \sum_{d=1}^{D} \mathbb{E}_{q_{\gamma_z}(\boldsymbol{z}|\boldsymbol{Y},\boldsymbol{X})} \left[\sum_{k=1}^{K} \log p_{\boldsymbol{\theta}_k} \left(\boldsymbol{y}_d \mid \boldsymbol{x}_d \right) \cdot \pi_{dk} \right] - D_{KL} \left(q_{\gamma_z}(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{Y}) \| p_{\psi_z}(\boldsymbol{z}) \right)$$

$$- D_{KL} \left(q_{\gamma_c}(\boldsymbol{C} \mid \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{Y}) \| p_{\psi_c}(\boldsymbol{C} \mid \boldsymbol{z}, \boldsymbol{X}) \right)$$

- For all samples in our dataset $\left(\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}\right), i \in [N]$

$$\max_{\phi,\gamma,\psi} \sum_{i=1}^{N} \mathcal{L}(\phi, \gamma, \psi; \boldsymbol{Y}^{(i)}, \boldsymbol{X}^{(i)})$$

Reconstruction

KL of the continuous latent variable

KL of the discrete latent variable

34

Experiments

Baselines

Model	Approach	Training Procedure	Generation	Reconstruction, Imputation, Super Resolution
GASP (2021) [5]	GAN	Minimax	Forward Pass	❌

35

Experiments

Baselines

Model	Approach	Training Procedure	Generation	Reconstruction, Imputation, Super Resolution
GASP (2021) [5]	GAN	Minimax	Forward Pass	❌
Functa (2022) [6]	Flow-based	Bilevel optimization	+ Extra Generative Model	Optimization procedure(s) per sample 

36

Experiments

Baselines

Model	Approach	Training Procedure	Generation	Reconstruction, Imputation,
GASP (2021) [5]	GAN	Minimax	Forward Pass	
Functa (2022) [6]	Flow-based	Bilevel optimization	+ Extra Generative Model	Optimization procedure(s) per sample 

$$\min_{\phi} -\log p(\phi) + \lambda \sum_{i \in \mathcal{I}} \|f_{\phi}(\mathbf{x}_i) - \mathbf{f}_i\|_2^2$$

Experiments

Baselines

Model	Approach	Training Procedure	Generation	Reconstruction, Imputation, Super Resolution
GASP (2021) [5]	GAN	Minimax	Forward Pass	❌
Functa (2022) [6]	Flow-based	Bilevel optimization	+ Extra Generative Model	Optimization procedure(s) per sample
VaMoH (ours)	VAE-based	**Single optimization**	**Forward Pass**	**Forward pass**

VAMoH provides a probabilistic generative model that is efficient, robust, and expressive for modeling distribution over functions.

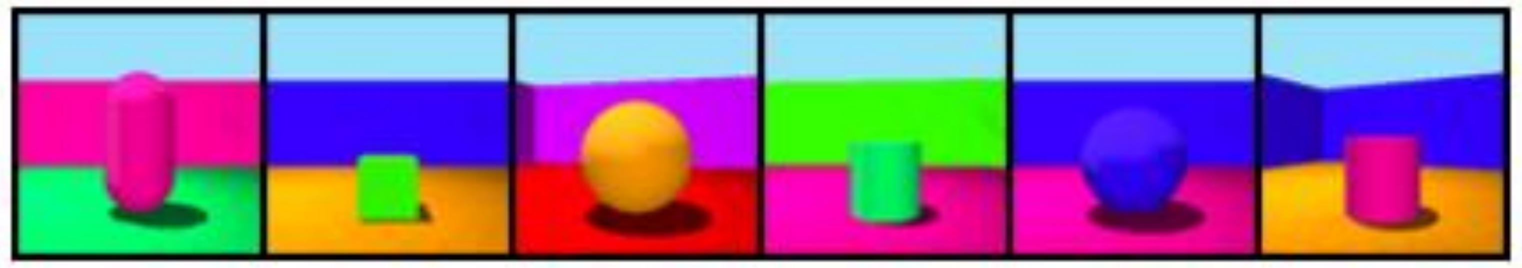
38

Experiments

Datasets

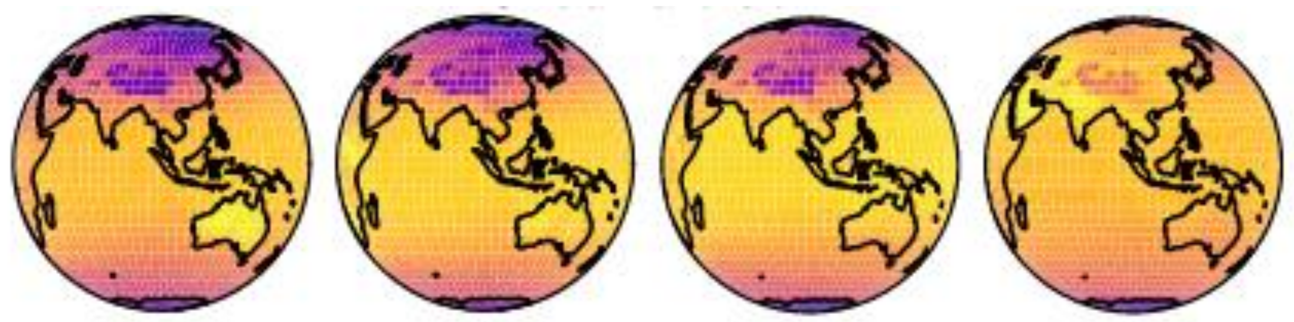
PolyMNIST (28x28)

Shapes3D (64x64)



CelebA-HQ (64x64)

ERA5 (Polar)

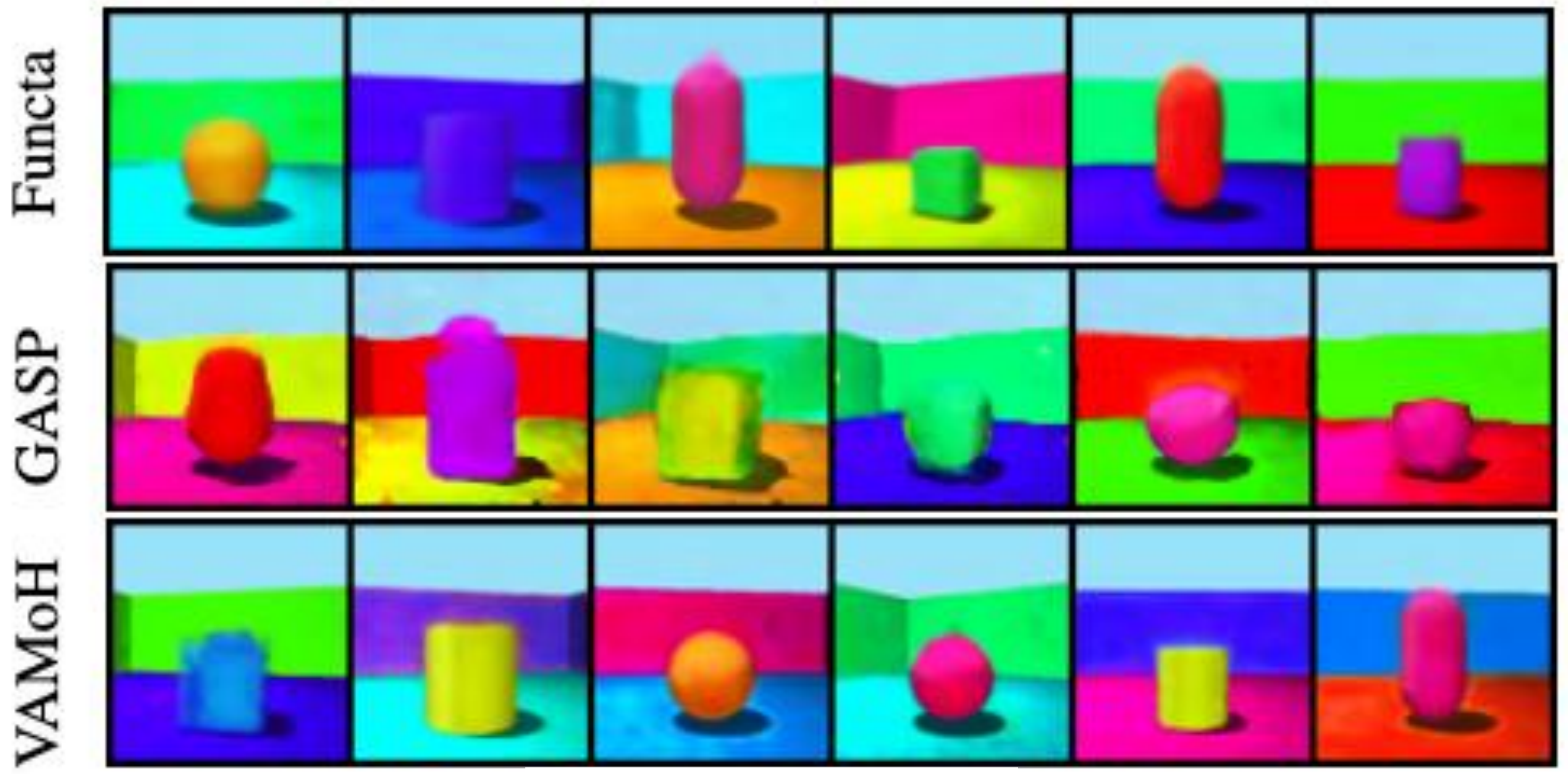


ShapeNET (Voxels)

Experiments

Generation

CelebA-HQ



Shapes3D

Experiments

Generation

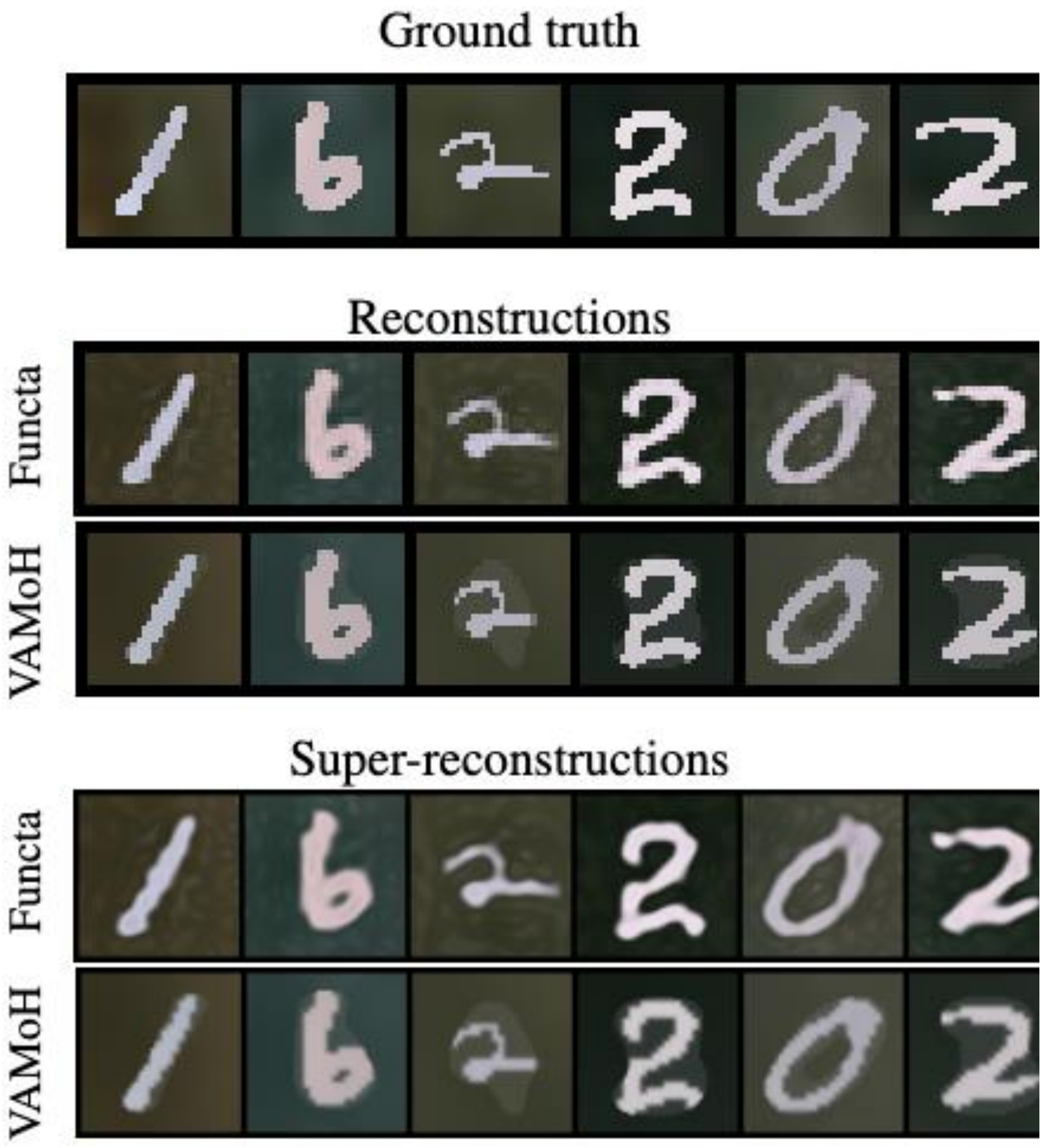
PolyMNIST

ERA5

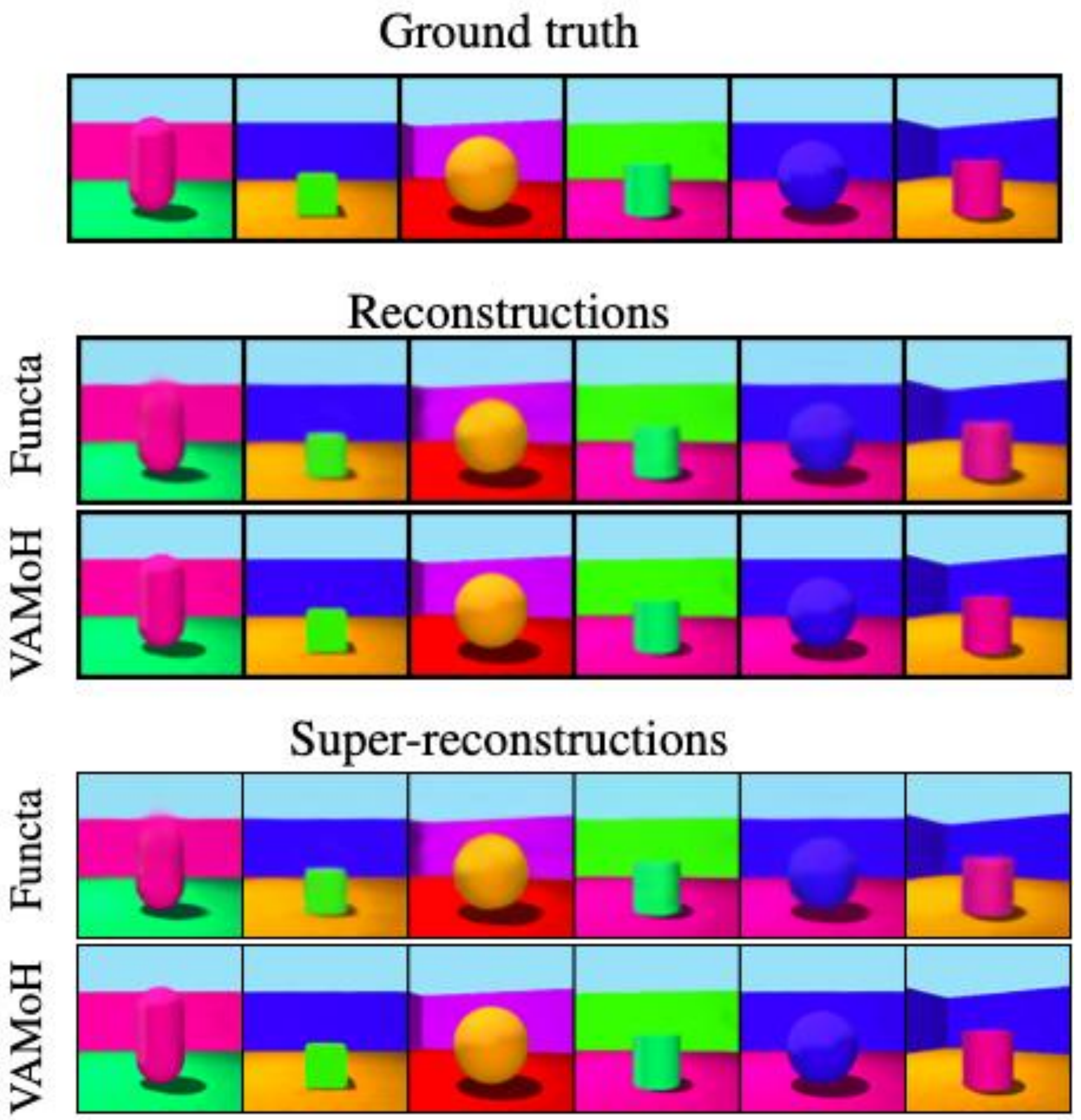
ShapeNET

Experiments

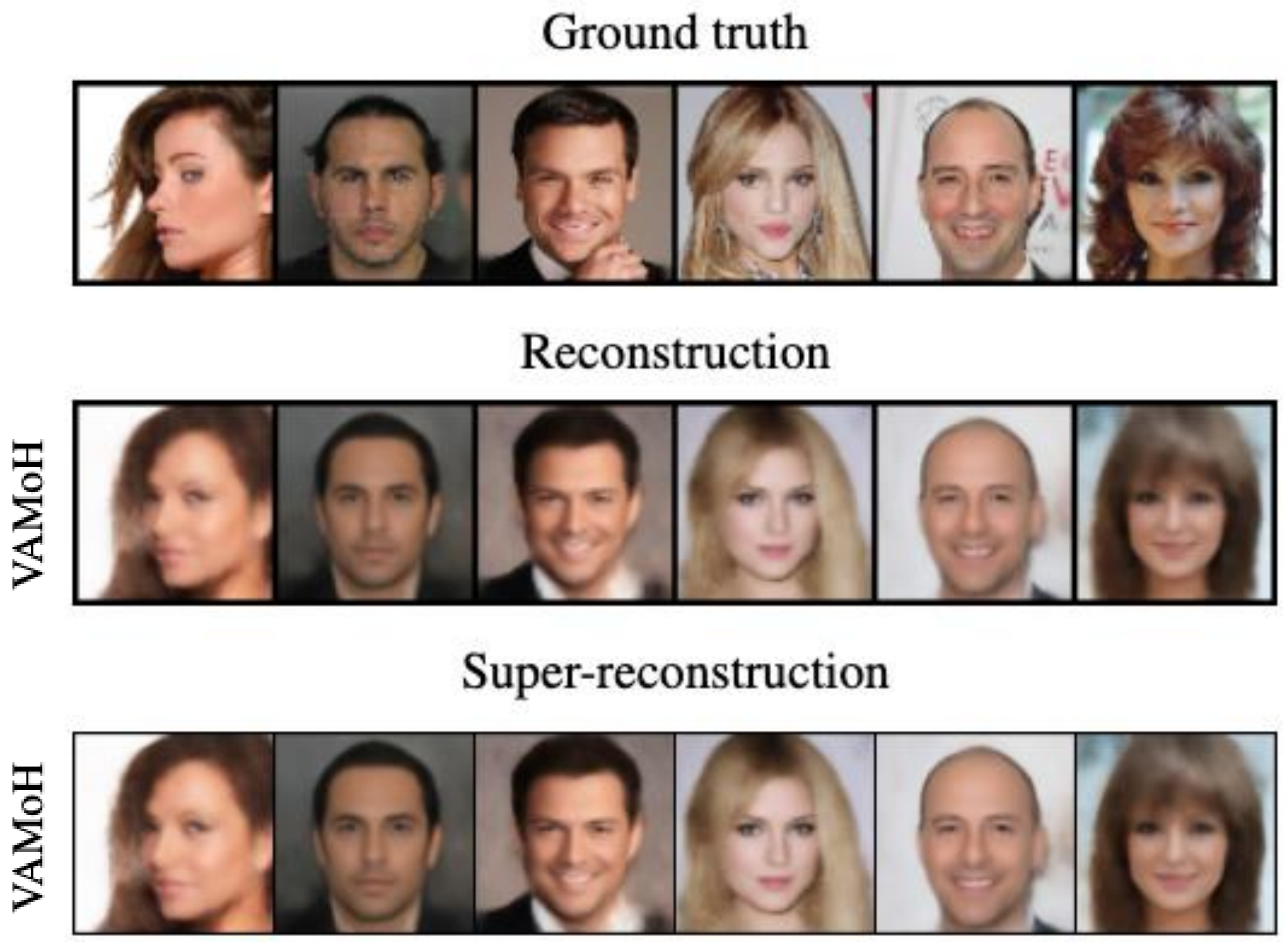
Reconstructions



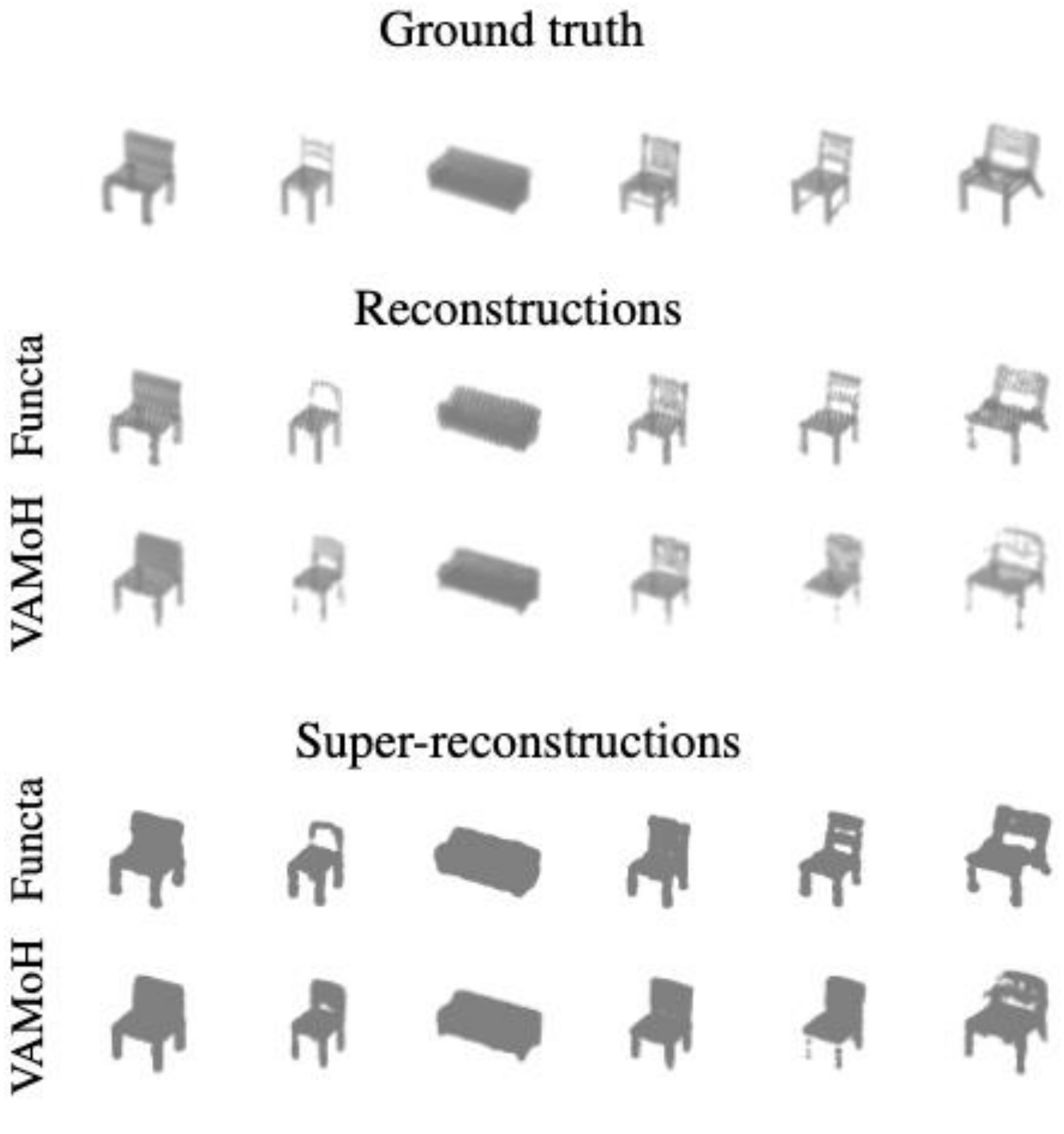
(c) POLYMNIST



SHAPES3D



CELEBA HQ



SHAPENET

Experiments

Inference times

Table 2: Comparison of inference time (seconds) for reconstruction task of VaMoH and Functa. On the right-most two columns, we show the speed improvement of VaMoH compared to Functa (3) which is trained with 3 gradient steps as suggested in the original paper [Dupont et al., 2022b] and Functa (10) which is trained with 10 gradient step to obtain the results of Functa depicted in Figures 16,17. Please note that these experiments are run on the same GPU device.

	Model Inference Time (secs)			Speed Improvement	
Dataset	VaMoH	Functa (3)	Functa (10)	vs. Functa (3)	vs. Functa (10)
POLYMNIST	**0.00453**	0.01648	0.05108	**x 3.64**	**x 11.28**
SHAPES3D	**0.00536**	0.01759	0.05480	**x 3.28**	**x 10.22**
CELEBA HQ	**0.00757**	0.01733	0.05381	**x 2.29**	**x 7.11**
ERA5	**0.00745**	0.01899	0.05932	**x 2.55**	**x 7.96**
SHAPENET	**0.00689**	0.02095	0.06576	**x 3.04**	**x 9.54**

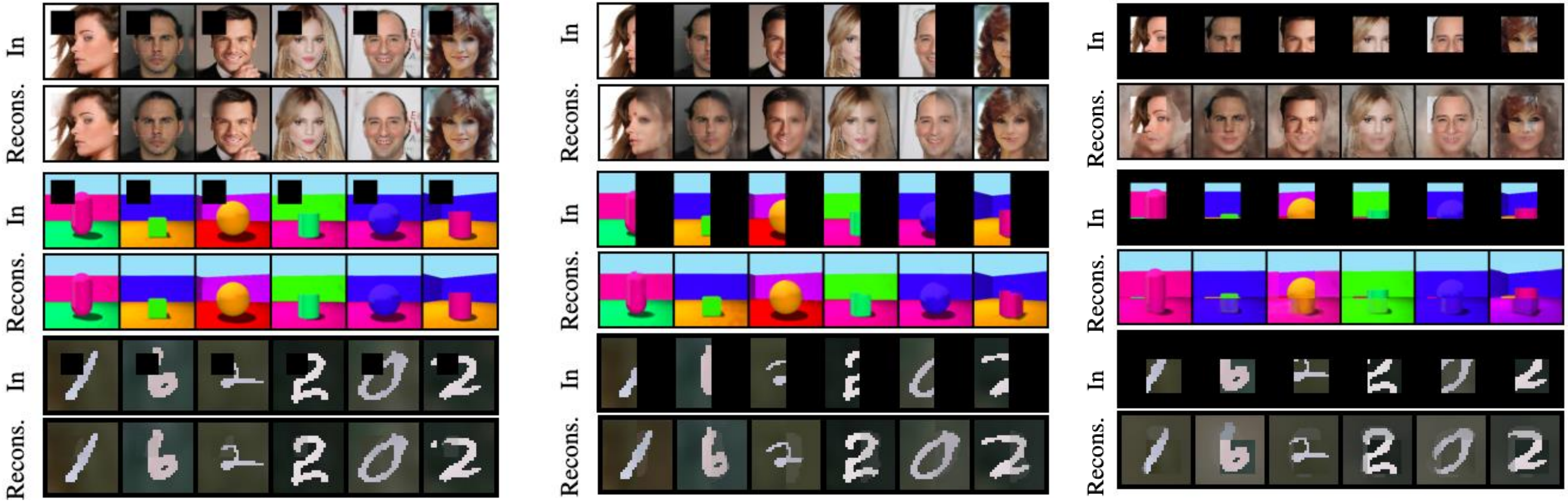
Reconstruction

	Model Inference Time (secs)			Speed Improvement	
Dataset	VaMoH	Functa (3)	Functa (10)	vs. Functa (3)	vs. Functa (10)
POLYMNIST	**0.00455**	0.01649	0.05109	**x 3.62**	**x 11.23**
SHAPES3D	**0.00544**	0.01768	0.05489	**x 3.25**	**x 10.09**
CELEBA HQ	**0.00833**	0.01729	0.05377	**x 2.08**	**x 6.46**
ERA5	**0.00790**	0.01997	0.06030	**x 2.53**	**x 7.63**
SHAPENET	**0.01440**	0.02089	0.06569	**x 1.45**	**x 4.56**

Super-reconstruction

Experiments

Image completion



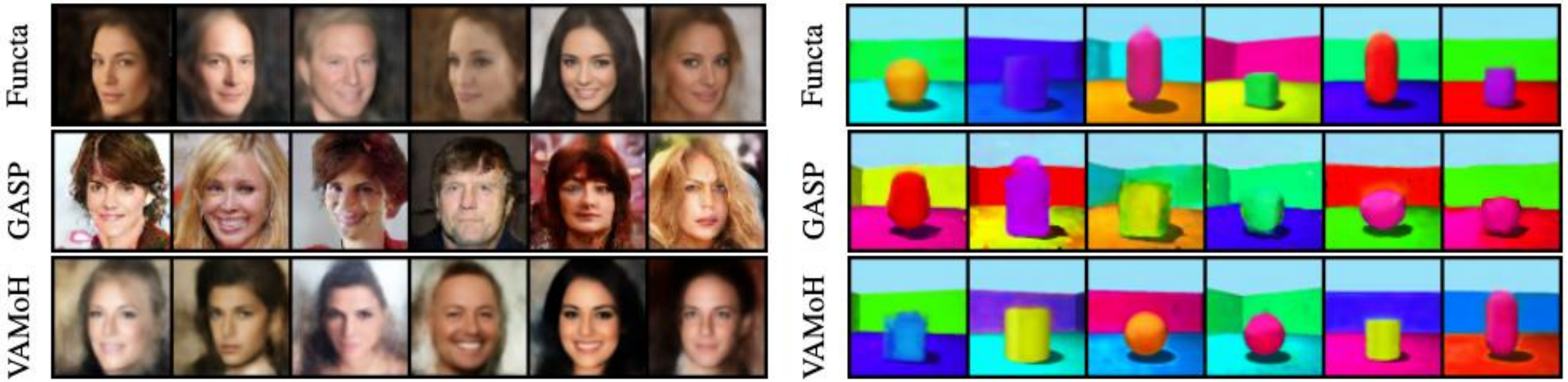
Missing a patch (in-painting) Missing half of the image Image out-painting

Proposed method (2)

Limitations of previous work

Flexibility of the latent space in [5, 6, 25]

- This makes generation quality poor.

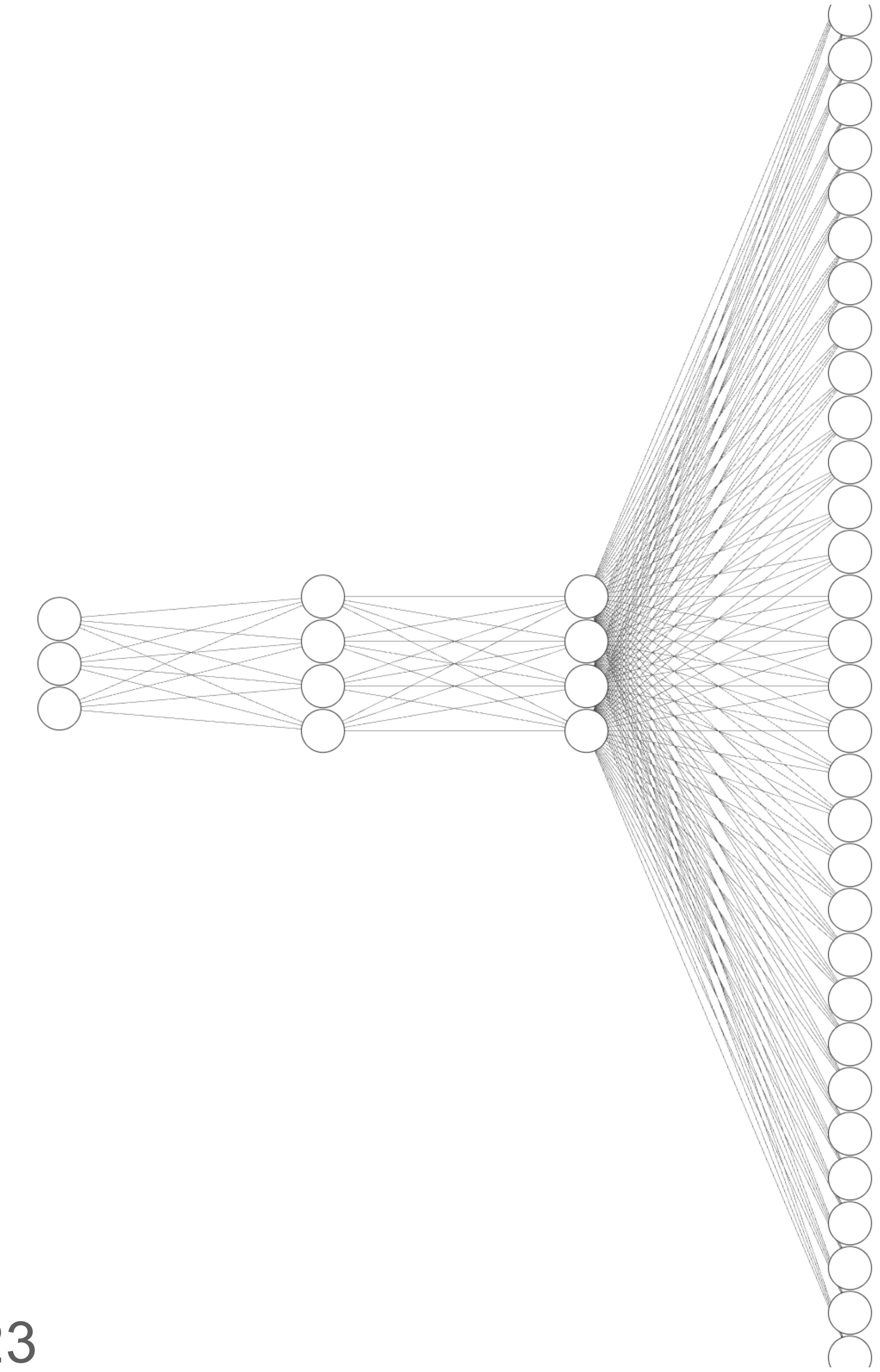


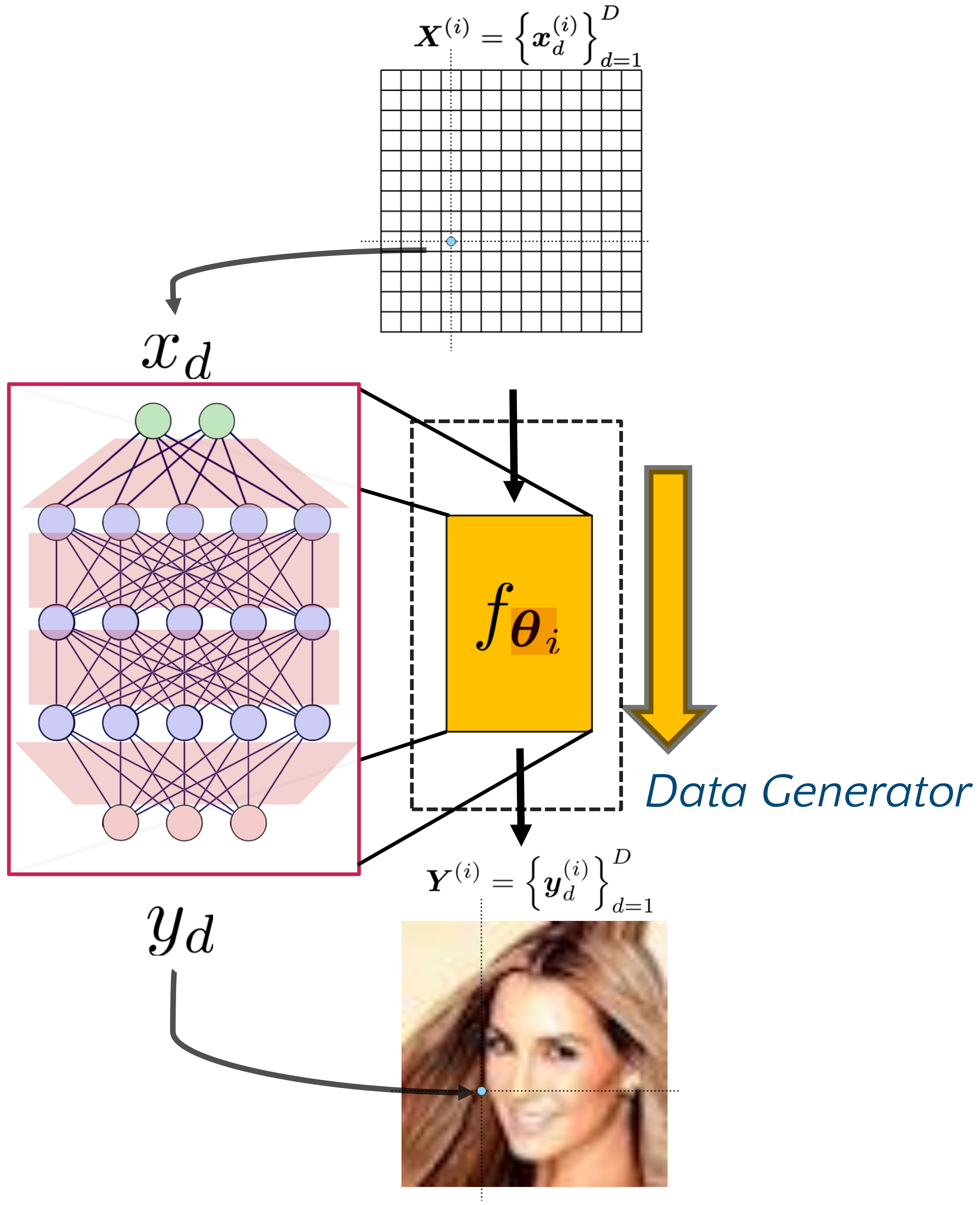
(a) CELEBA HQ (b) SHAPES3D

[5] Dupont et at., 2020 [6] Dupont et at., 2022 [25] Koyuncu et at., 2023

Limitations of previous work

Hypernetwork bottleneck in [5, 25]



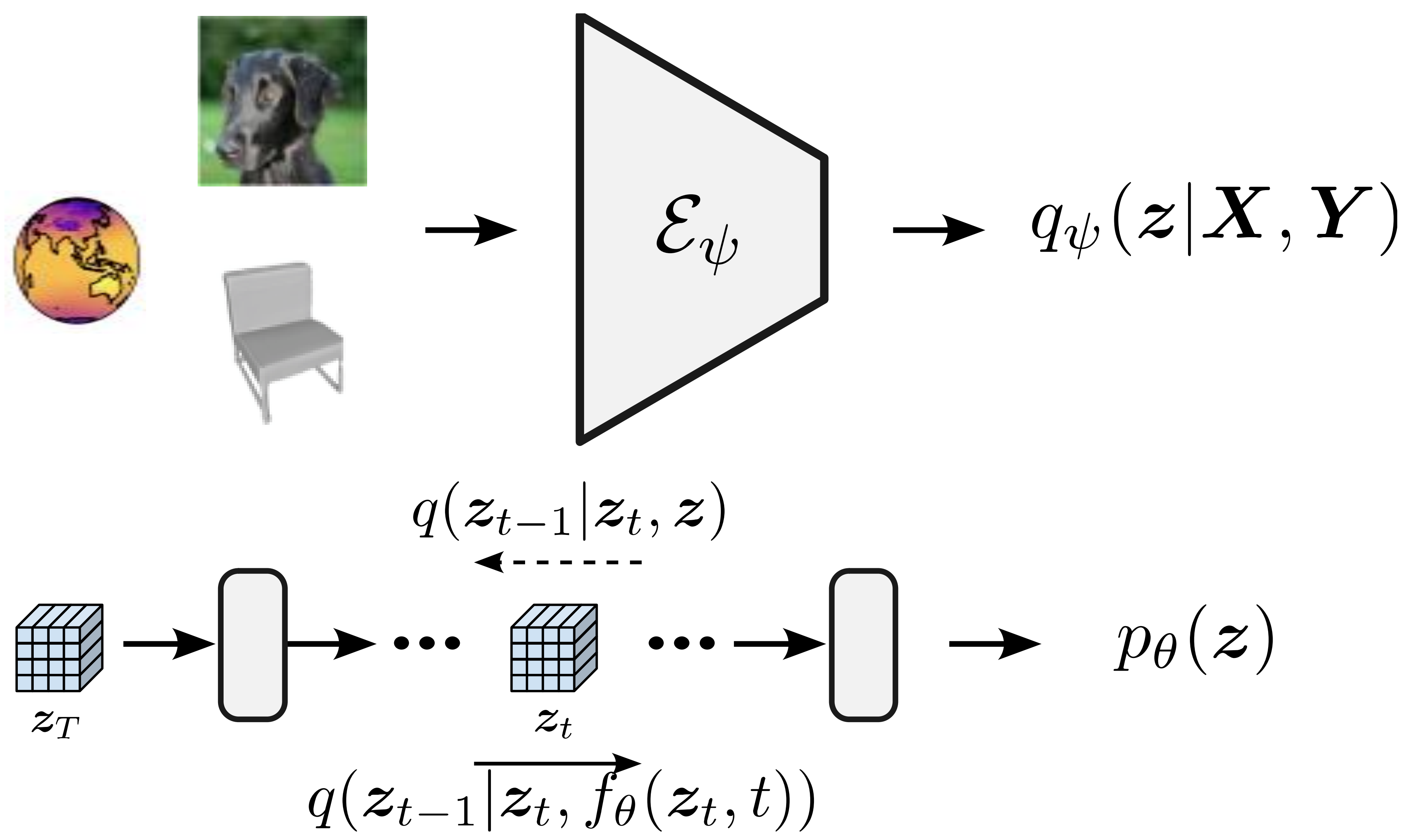


[5] Dupont et at., 2020 [25] Koyuncu et at., 2023

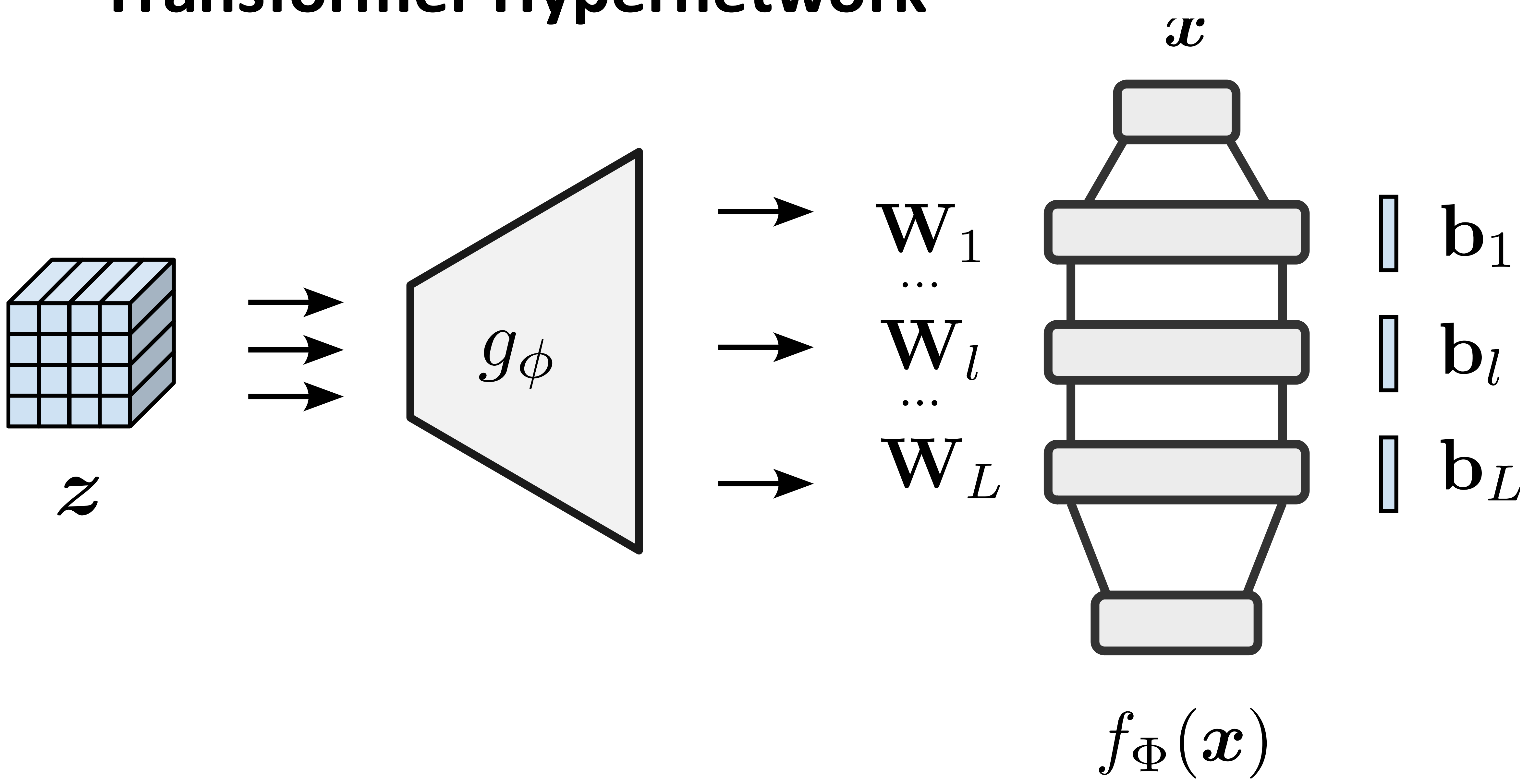
Proposed methods (2)

Hyper-Transforming Latent Variable Models [27] (LDMI)

Latent Diffusion [28]



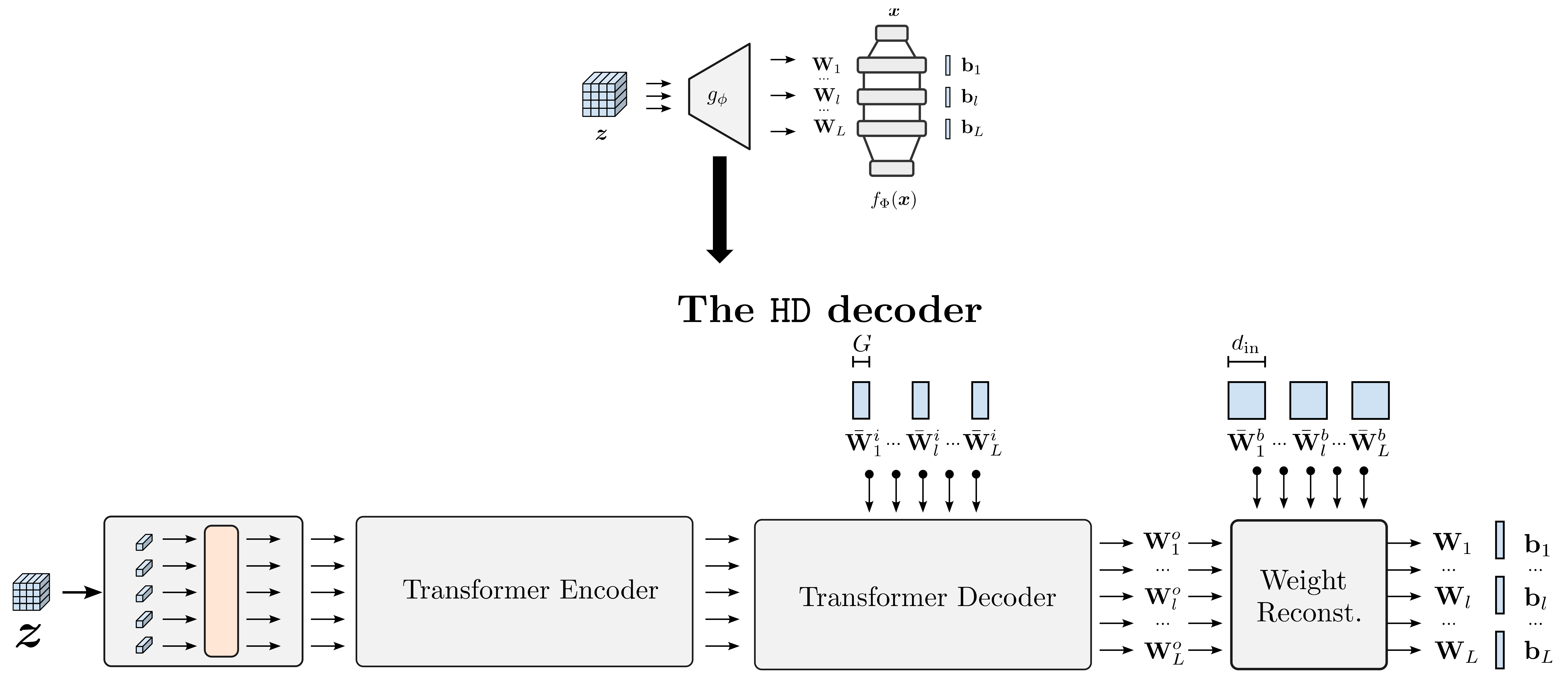
Transformer Hypernetwork



[27] Peis et at., 2025

Proposed methods (2)

The HD decoder



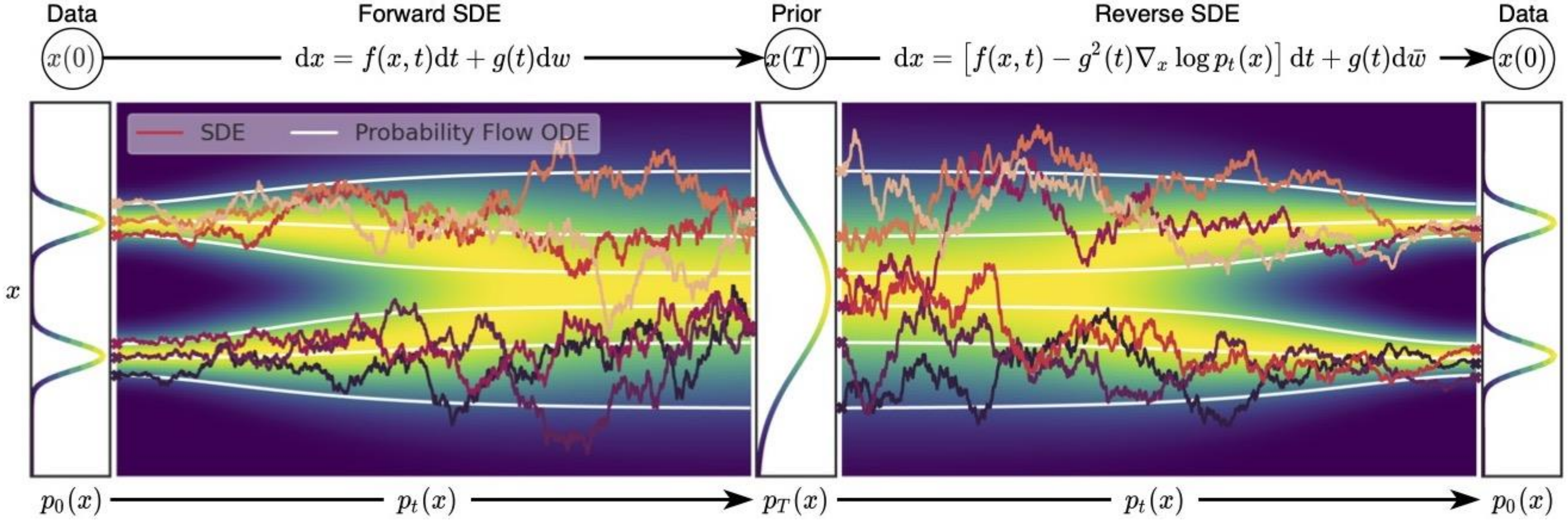
[27] Peis et at., 2025

LDMI

Diffusion Models [29]

Denoising Score Matching

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_t \left\{ \lambda(t) \mathbb{E}_{\mathbf{x}(0)} \mathbb{E}_{\mathbf{x}(t)|\mathbf{x}(0)} \left[\left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log p_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0)) \right\|_2^2 \right] \right\}$$

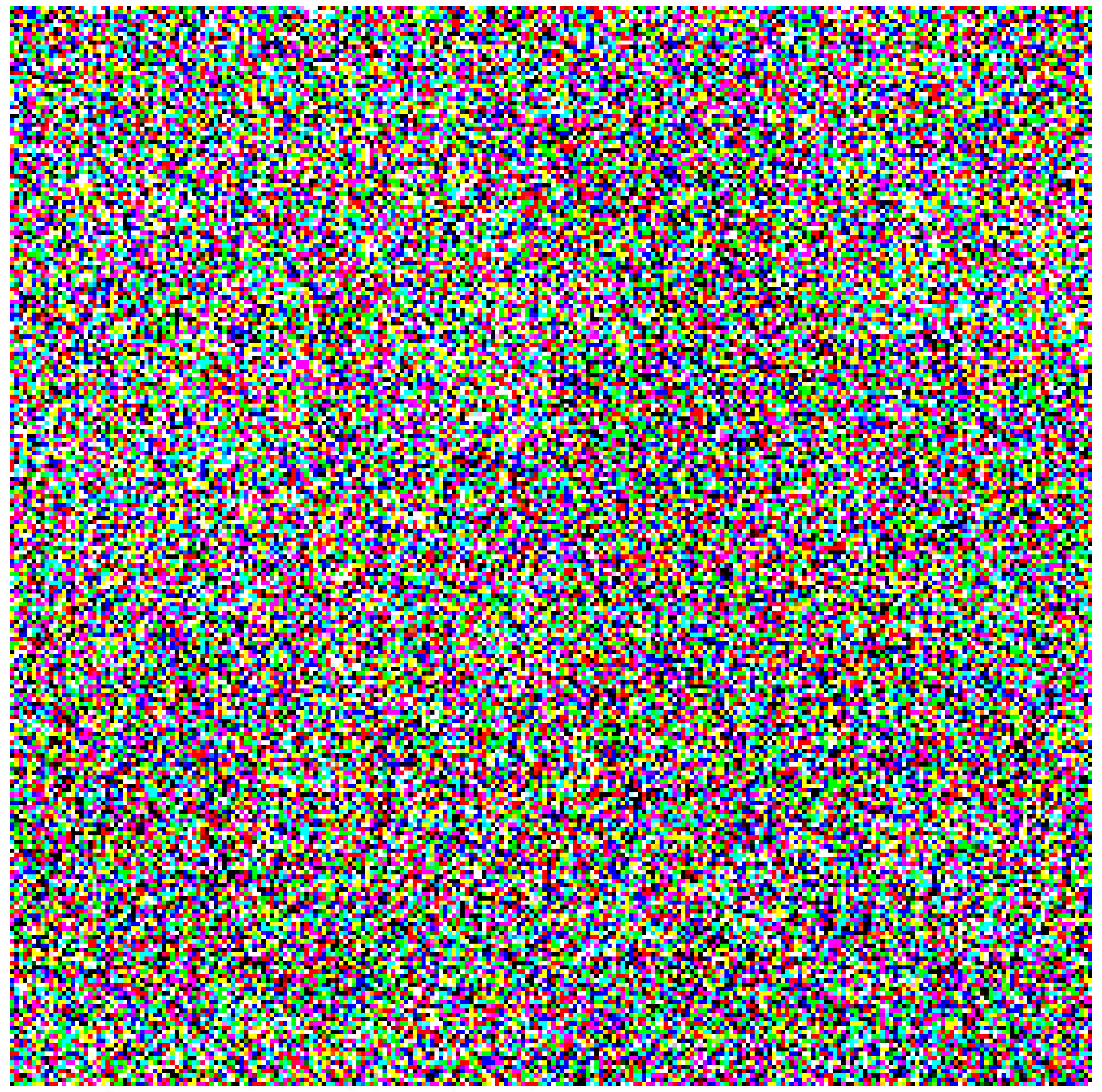


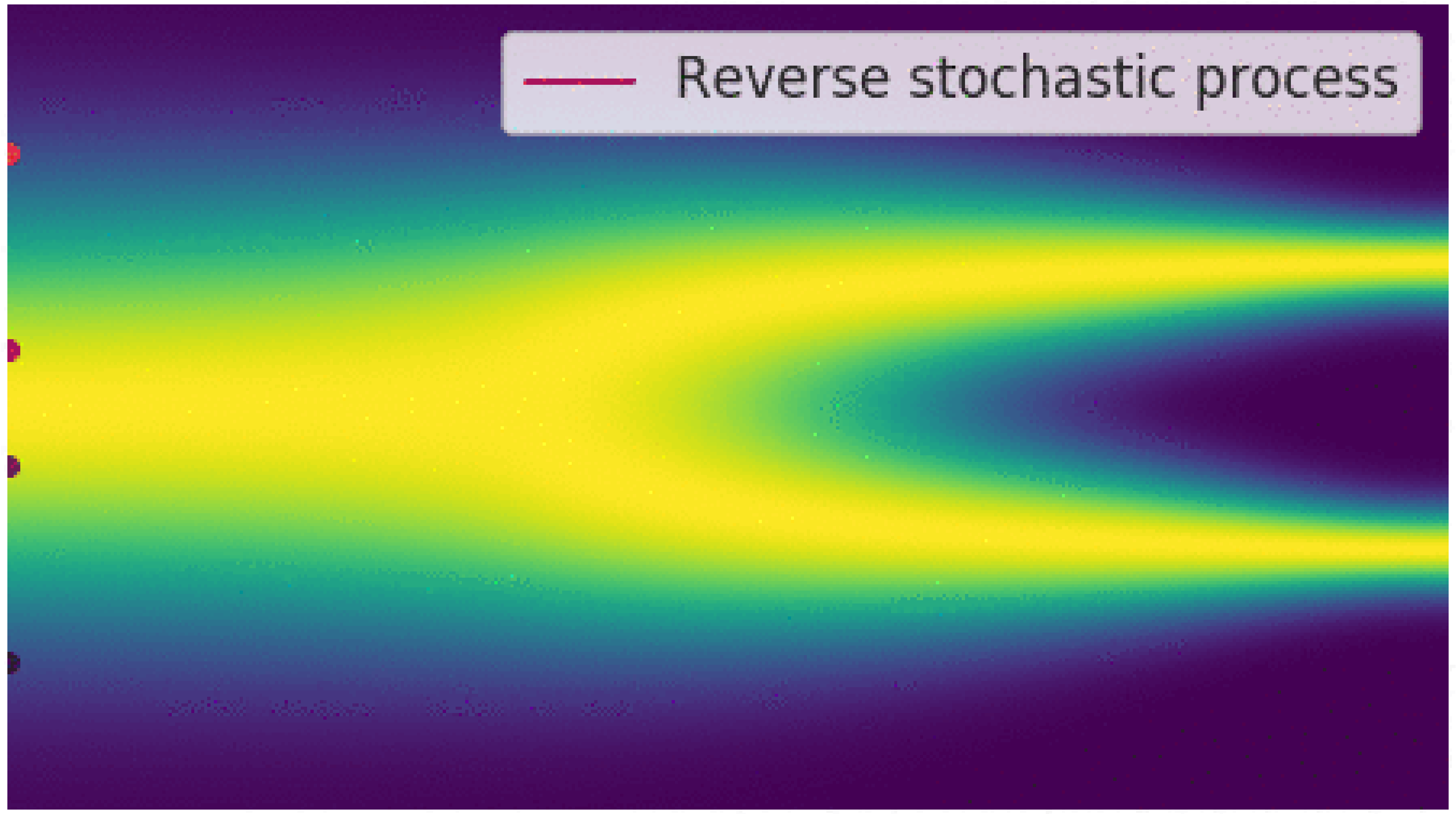
[29] Song et at., 2020

LDMI

Diffusion Models [29]

$$s_\theta(\boldsymbol{x}_t, t)$$





[29] Song et at., 2020

LDMI

DDPM [30]

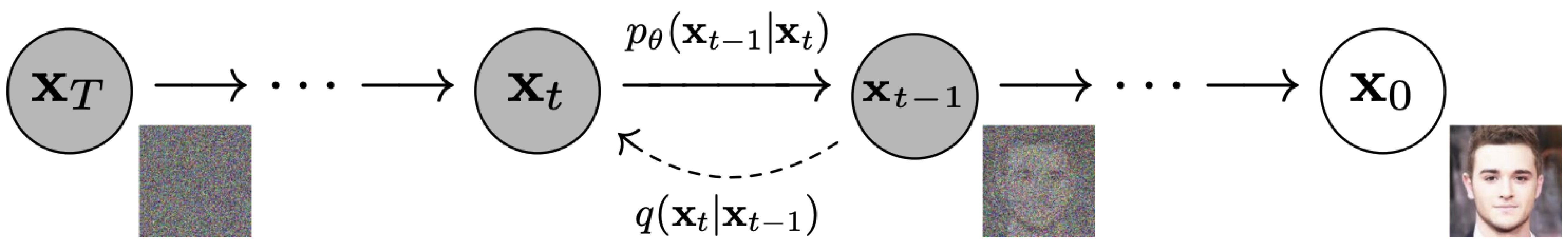


Figure 2: The directed graphical model considered in this work.

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

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right), \quad q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$$

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right) \qquad \alpha_t := 1-\beta_t \qquad \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$$

$$\mathbb{E}_q\Big[\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right)}_{L_T} + \sum_{t>1} \underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)}_{L_{t-1}} \underbrace{- \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}_{L_0}\Big]$$

[30] Ho et at., 2020

LDMI

DDPM [30]

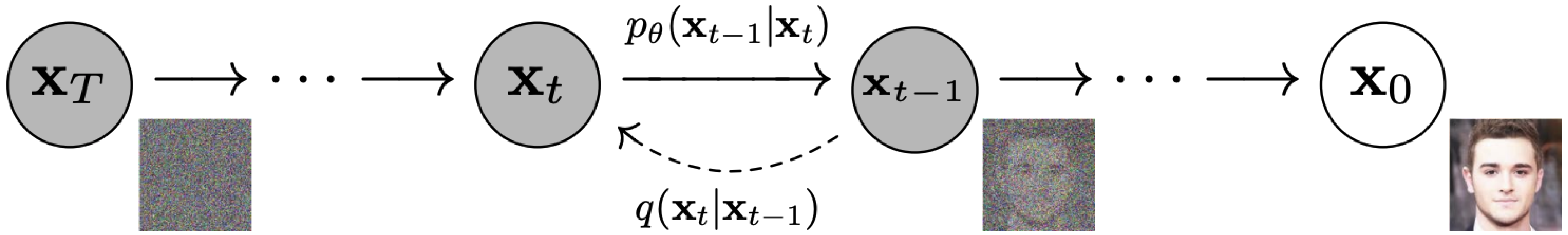


Figure 2: The directed graphical model considered in this work.

$$\mathbb{E}_q[\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right)}_{L_T}+\sum_{t>1} \underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)}_{L_{t-1}} \underbrace{-\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}_{L_0}]$$

$$\Downarrow$$

$$\mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2 \sigma_t^2 \alpha_t\left(1-\bar{\alpha}_t\right)}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0+\sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}, t\right)\right\|^2\right]$$

LDMI

DDPM [30]

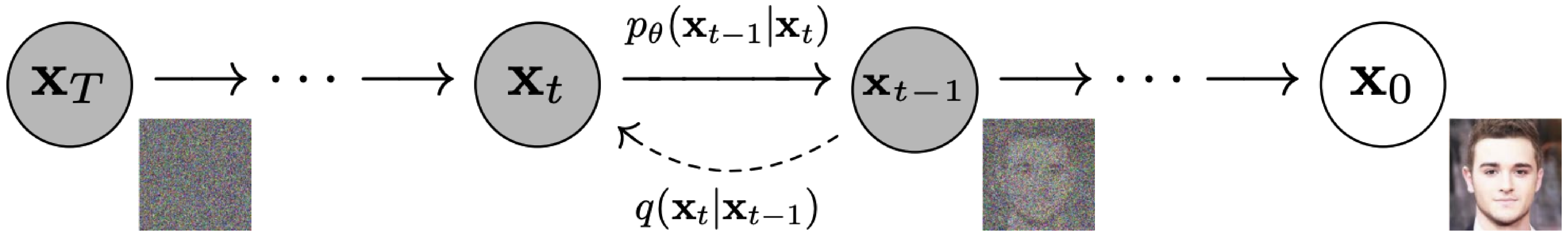


Figure 2: The directed graphical model considered in this work.

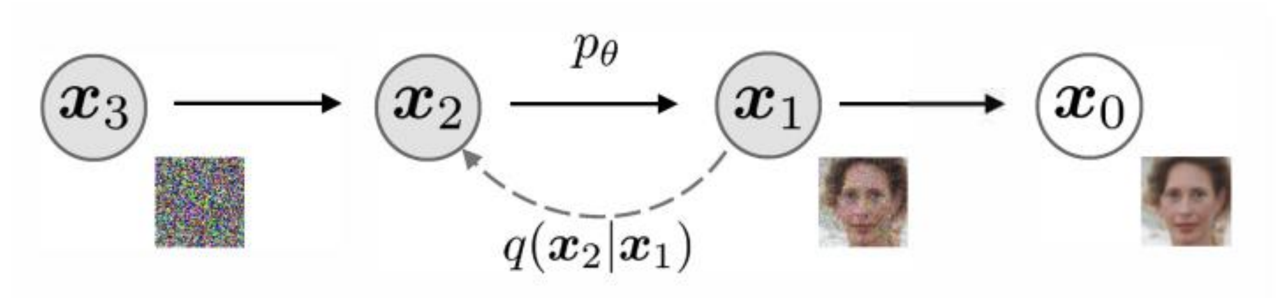
$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t\left(1-\bar{\alpha}_t\right)}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t\right)\right\|^2\right]$$

$$L_{\mathrm{simple}}\left(\theta\right):=\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t\right)\right\|^2\right]$$

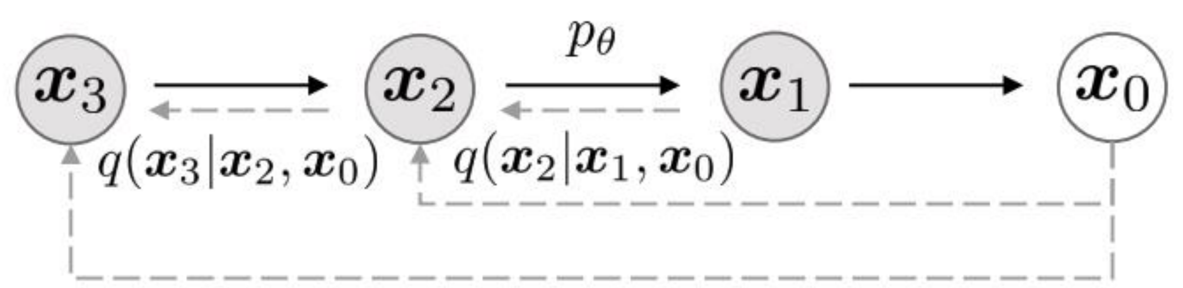
LDMI

DDIM [31]

Markovian (DDPM)



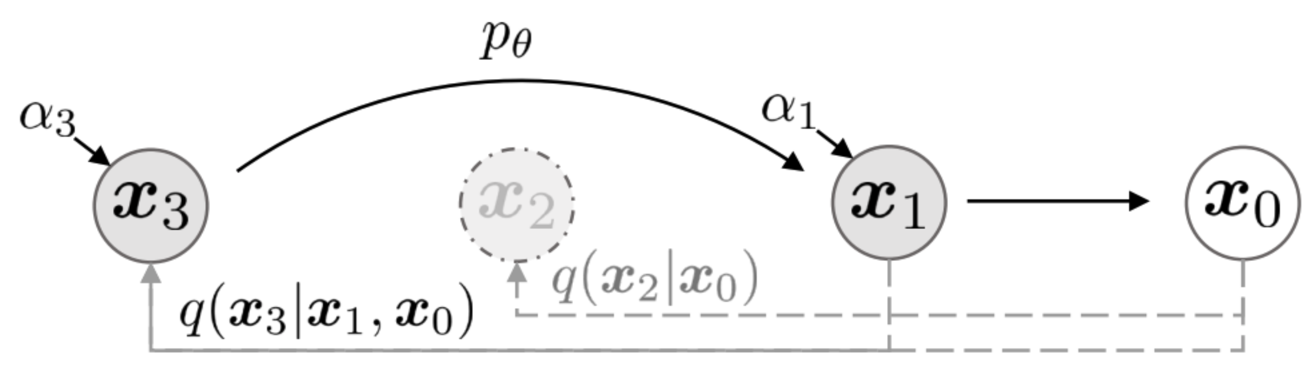
Non-Markovian (DDIM)



- Define a Non-Markovian Inference Model.
- The objective is the same!

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right]$$

- Using the same model, you can sample in fewer steps!



LDMI

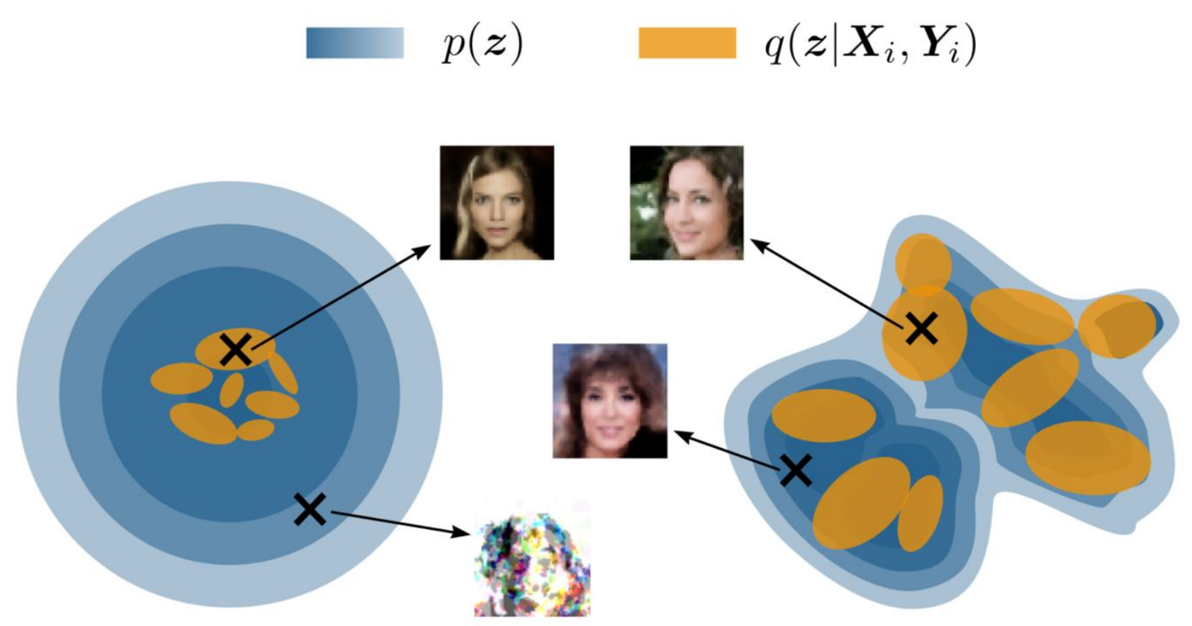
Latent Diffusion Models [28]

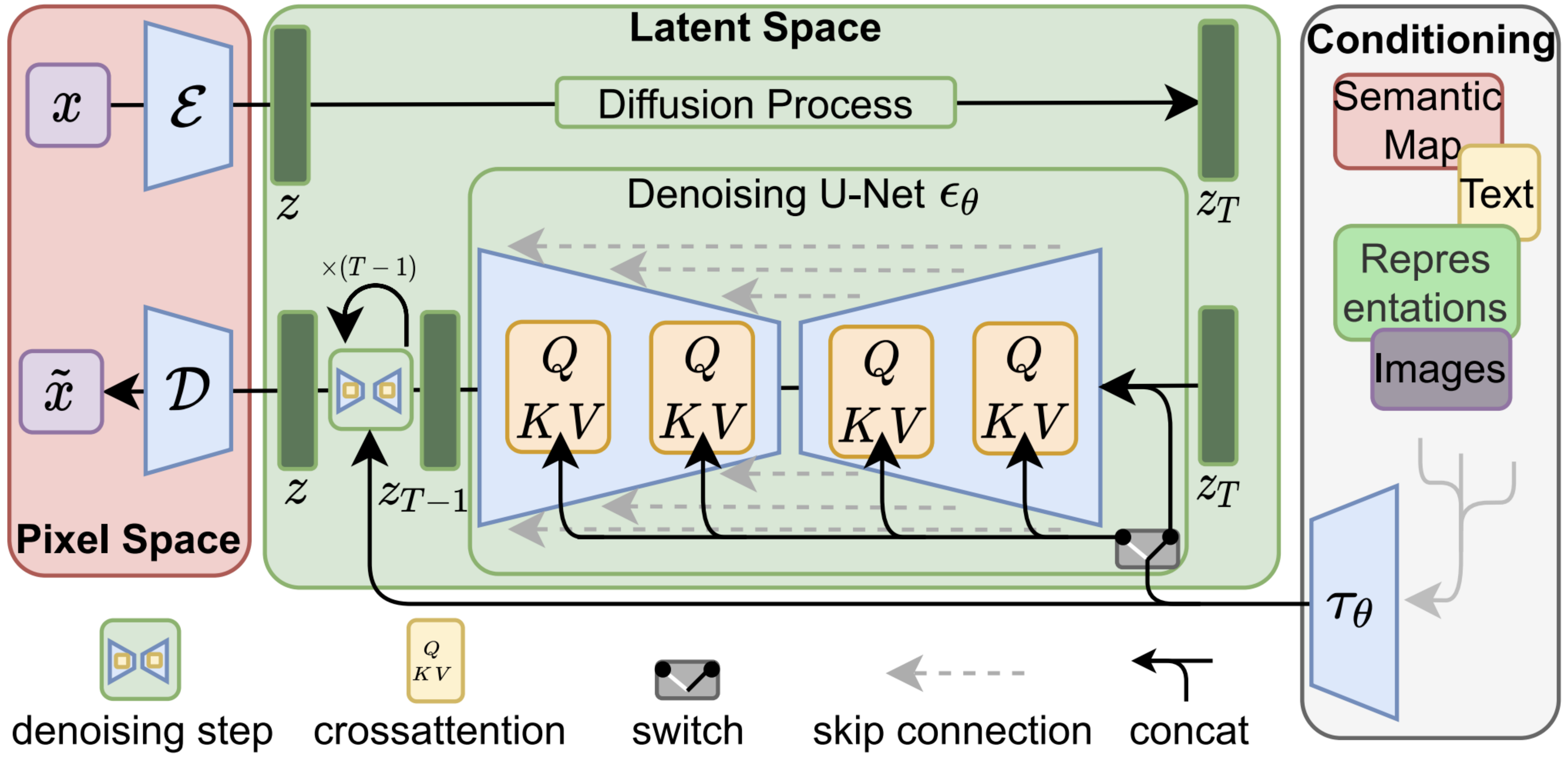
- First stage:

$$\mathcal{L}_{\text{VAE}}(\phi,\psi) = \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{X})}\left[\log p_\Phi(\boldsymbol{X})\right] - \beta \cdot D_{\text{KL}}\left(q_\psi(\boldsymbol{z} \mid \boldsymbol{X}) \| p(\boldsymbol{z})\right), + \mathcal{L}_{\text{perceptual}} + \mathcal{L}_{\text{GAN}}$$

- Second stage:

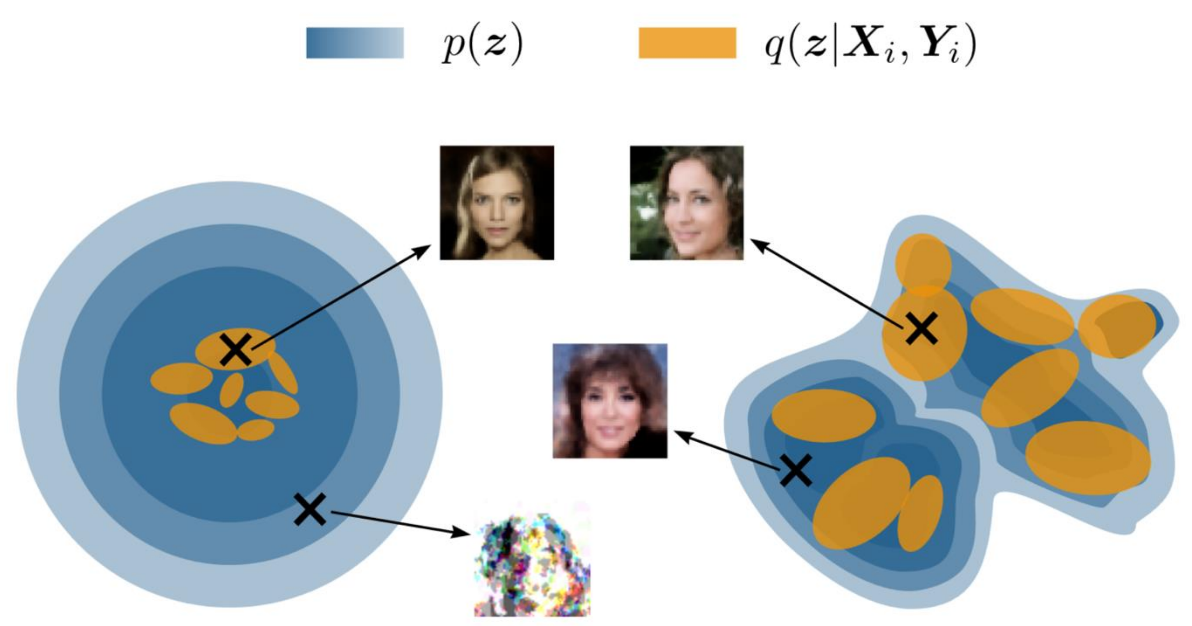
$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{z},\epsilon,t}\left[\lambda(t)\left\|\epsilon - \epsilon_\theta\left(\boldsymbol{z}_t, t\right)\right\|^2\right],$$

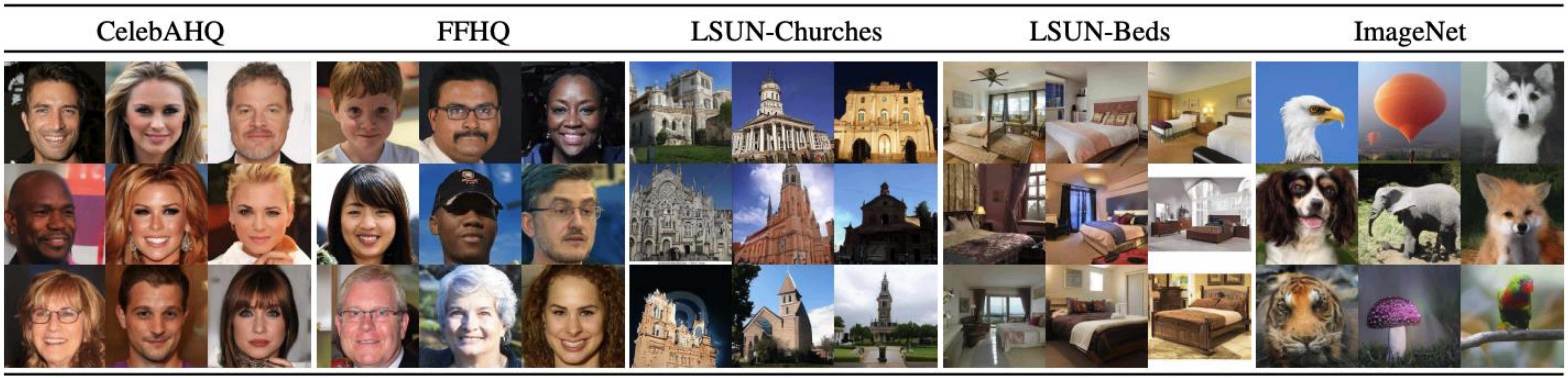




LDMI

Latent Diffusion Models





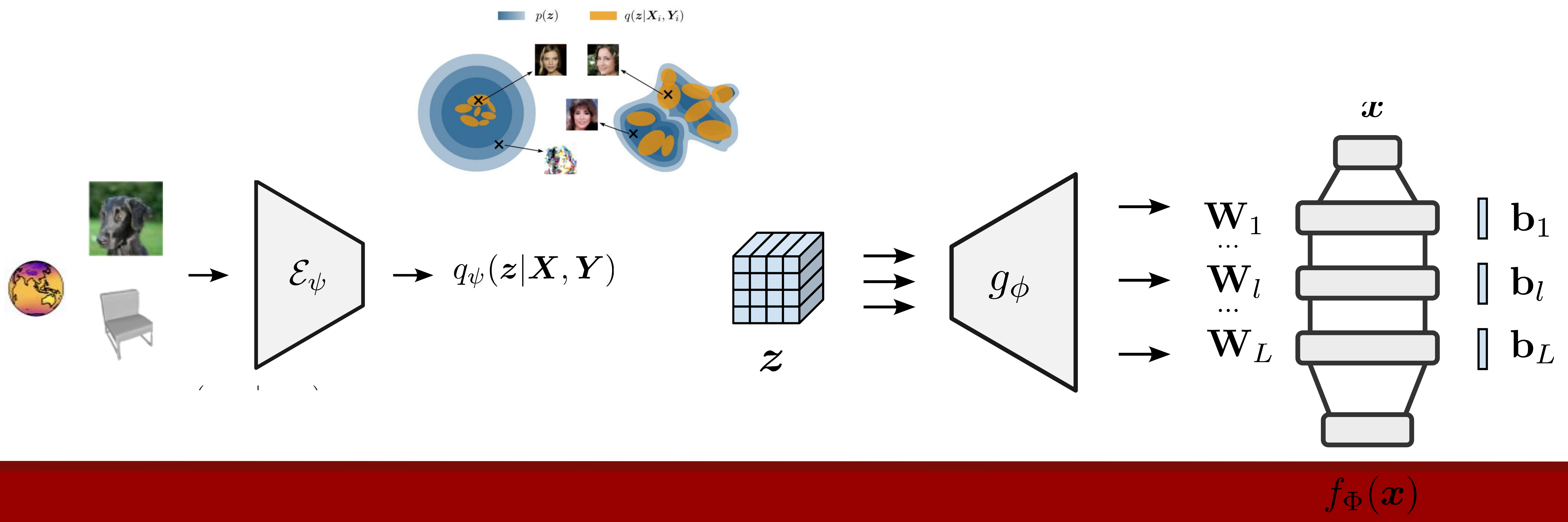
LDMI

Latent Diffusion Models for Implicit Neural Representations

1. We will train an *“under-regularized”* autoencoder (VAE or VQ-VAE) to accurately represent data in a (tensor-shaped) latent space.

   - The latents are mapped into INRs using our **transformer–based hypernetwork decoder**.

$$\mathcal{L}_{\text{VAE}}(\phi,\psi) = \mathbb{E}_{q_\psi(\boldsymbol{z}\mid\boldsymbol{X},\boldsymbol{Y})}\left[\log p_\Phi(\boldsymbol{Y}\mid\boldsymbol{X})\right] - \beta\cdot D_{\text{KL}}\left(q_\psi(\boldsymbol{z}\mid\boldsymbol{X},\boldsymbol{Y})\|p(\boldsymbol{z})\right),$$

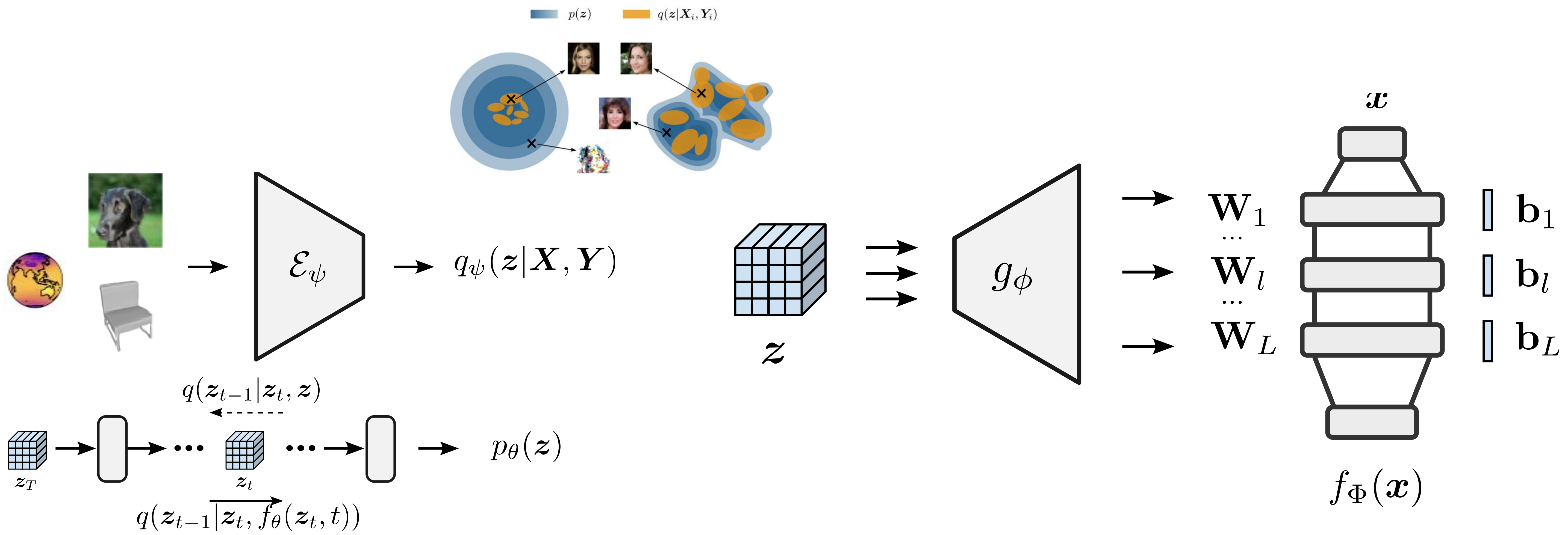


LDMI

Latent Diffusion Models for Implicit Neural Representations

2. We will fit a Diffusion Model (DDPM) to the learned latent space.

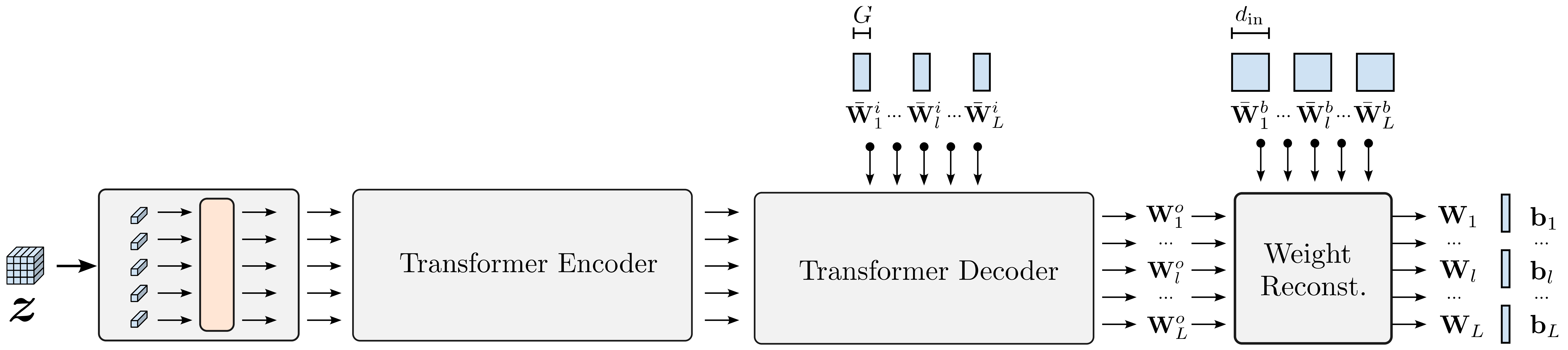
$$\mathcal{L}_{\mathrm{DDPM}} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{z},\epsilon,t}\left[\lambda(t)\left\|\epsilon - \epsilon_\theta\left(\boldsymbol{z}_t, t\right)\right\|^2\right],$$



LDMI

The Hyper-Transformer Decoder

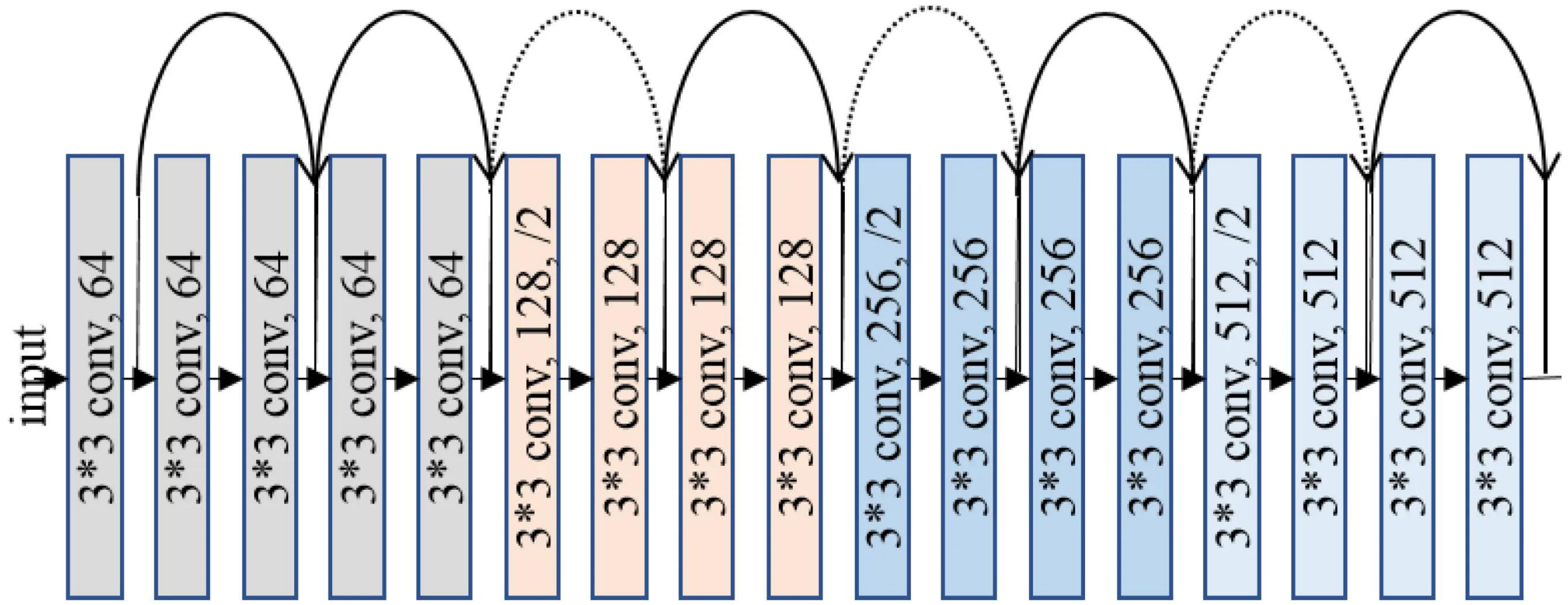
- The latents are **tokenized** (following ViT [32]).
- Two sets of globally shared, learnable parameters:
  - Compressed weights that cross-attend the latent tokens.
  - Full weights to expand the compressed weights.



LDMI

ResNet encoders

- The data is stored in a structured representation.
- We can make use of powerful encoders tailored to structured data.

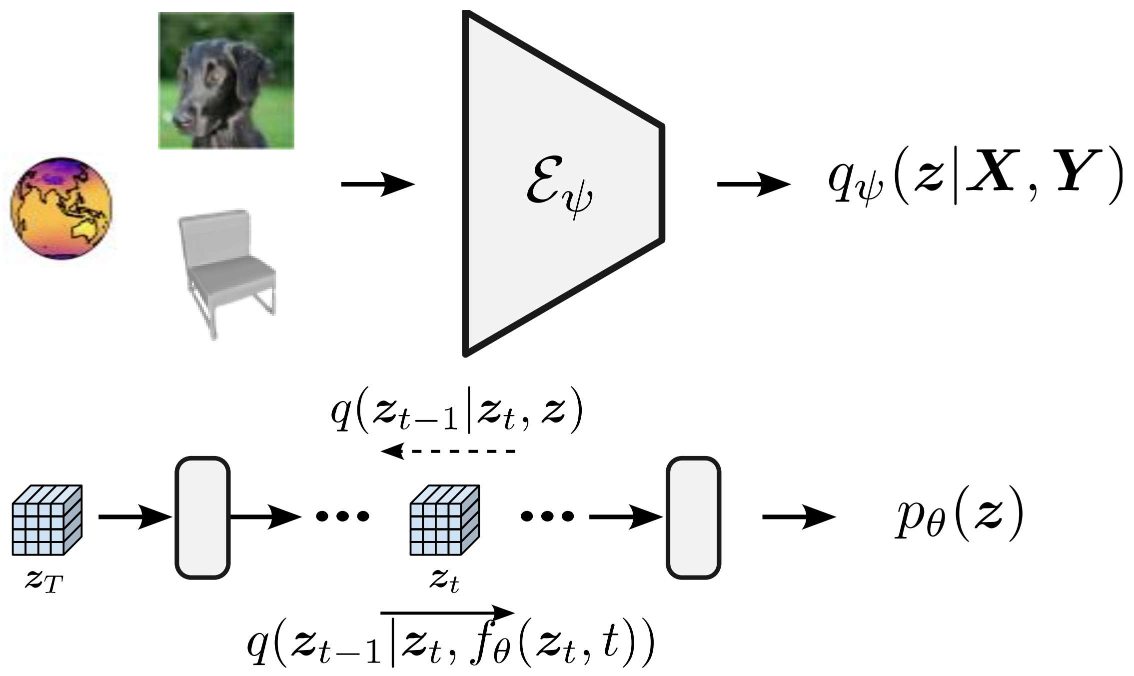


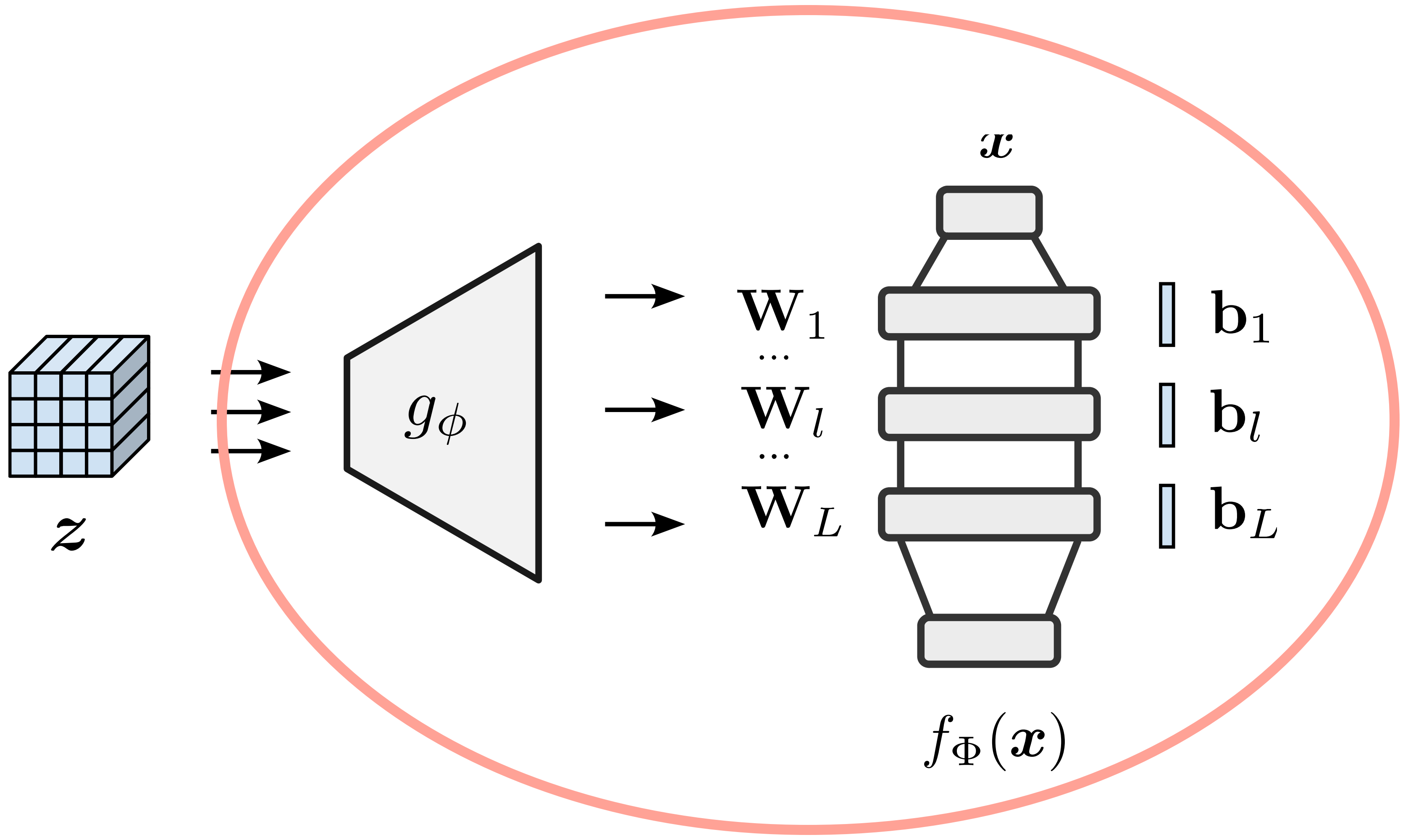
LDMI

Hyper-Transforming

- We can download pre-trained LDMs and just re-train only our decoder!

$$\mathcal{L}_{\text{HT}}(\phi) = \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{X}_m, \boldsymbol{Y}_m)} \left[\log p_\Phi(\boldsymbol{Y} \mid \boldsymbol{X})\right] + \mathcal{L}_{\text{perceptual}} + \mathcal{L}_{\text{GAN}}$$





Pretrained LDMs

Datset	Task	Model	FID	IS	Prec	Recall	
CelebA-HQ	Unconditional Image Synthesis	LDM-VQ-4 (200 DDIM steps, eta=0)	5.11 (5.11)	3.29	0.72	0.49	https://omr diffusion/c
FFHQ	Unconditional Image Synthesis	LDM-VQ-4 (200 DDIM steps, eta=1)	4.98 (4.98)	4.50 (4.50)	0.73	0.50	https://omr diffusion/ff
LSUN-Churches	Unconditional Image Synthesis	LDM-KL-8 (400 DDIM steps, eta=0)	4.02 (4.02)	2.72	0.64	0.52	https://omr diffusion/ls

Experiments

Datasets

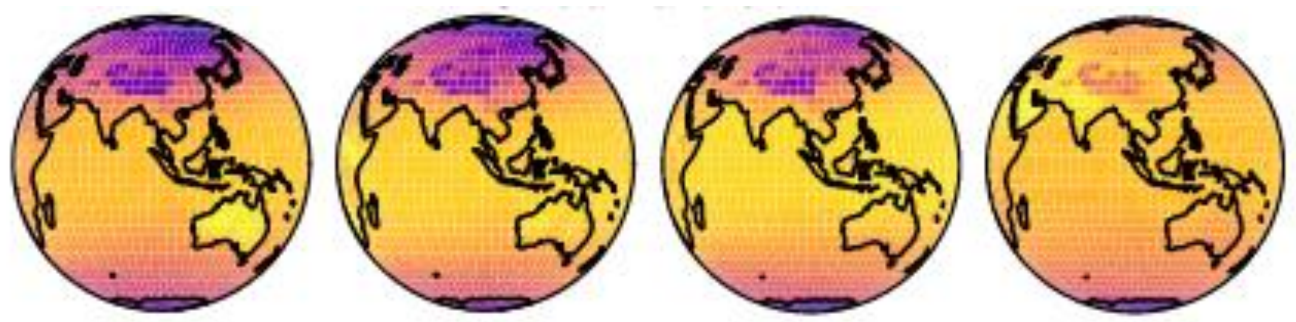
CelebA (64x64)

CelebA-HQ (64x64)

CelebA-HQ (256x256)

ImageNet (256x256)

ERA5 (Polar)



ShapeNET (Voxels)

Experiments

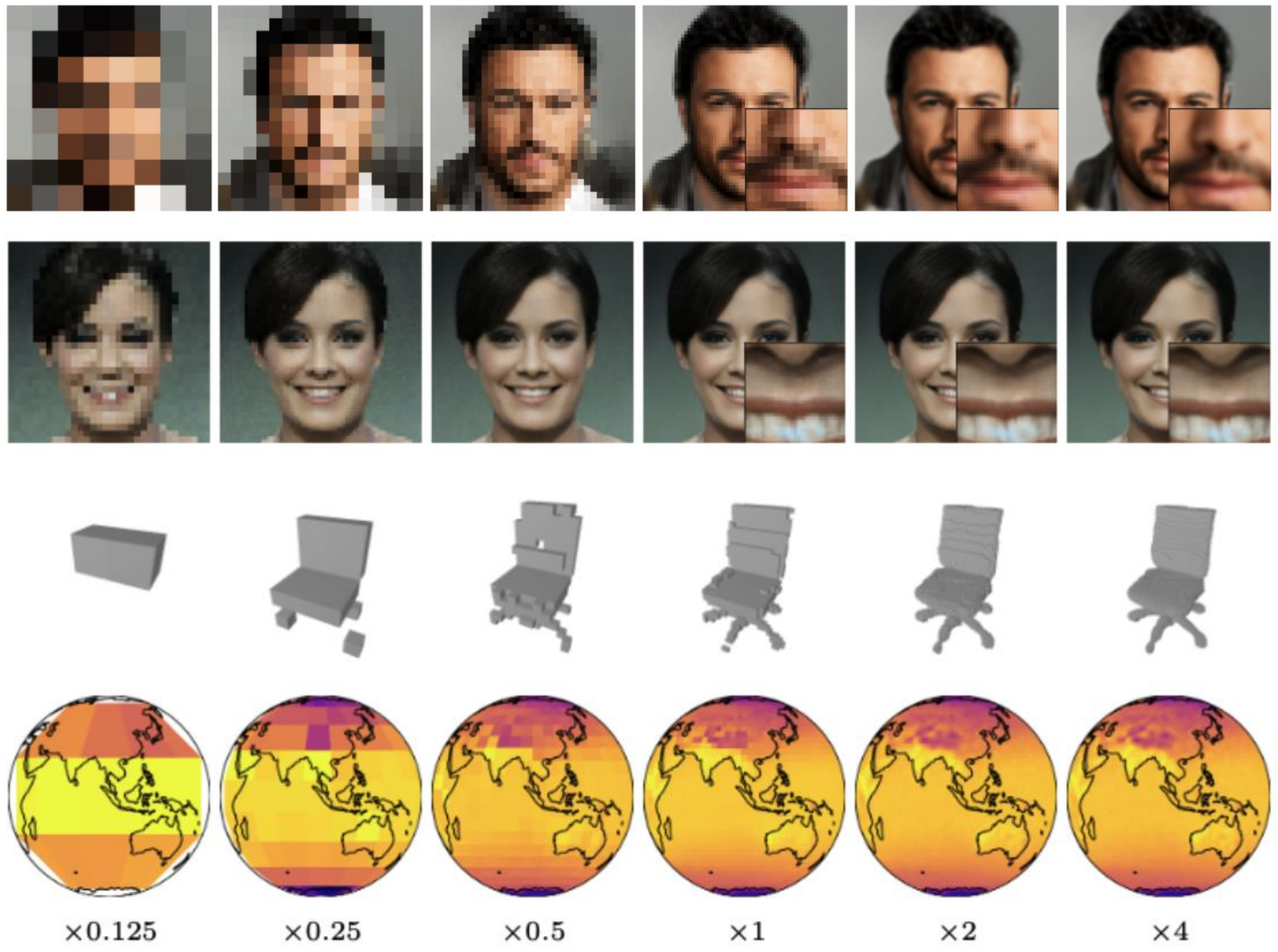
Baselines

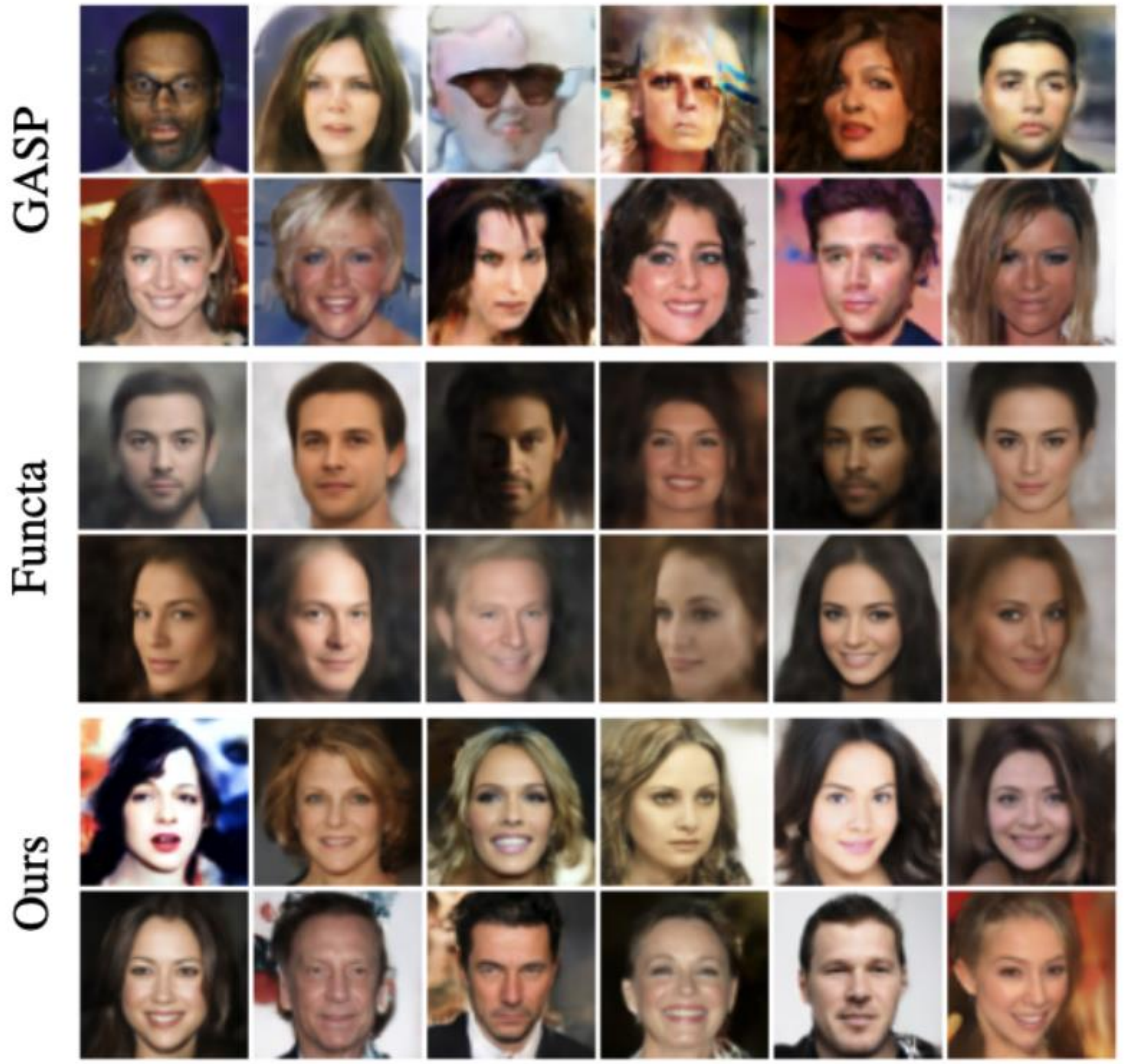
Model	Approach	Training Procedure	Generation	Reconstruction, Imputation, Super Resolution	Scalable	Flexible
GASP (2021) [5]	GAN	Minimax	Forward Pass	✗	✗	✗
Functa (2022) [6]	Flow-based	Bilevel optimization	+ Extra Generative Model	Optimization procedure(s) per sample	✗	✗
VaMoH (ours)	VAE-based	**Single optimization**	**Forward Pass**	**Forward pass**	✗	✗
LDMI (ours)	LDM-based	**Hyper-Transforming**	**Forward Pass**	**Forward pass**	✓	✓

LDMI enhances efficiency, scalability quality of the learned representations.

Experiments

Generation: qualitative results





(a) CelebA-HQ

Experiments

Generation: quantitative results

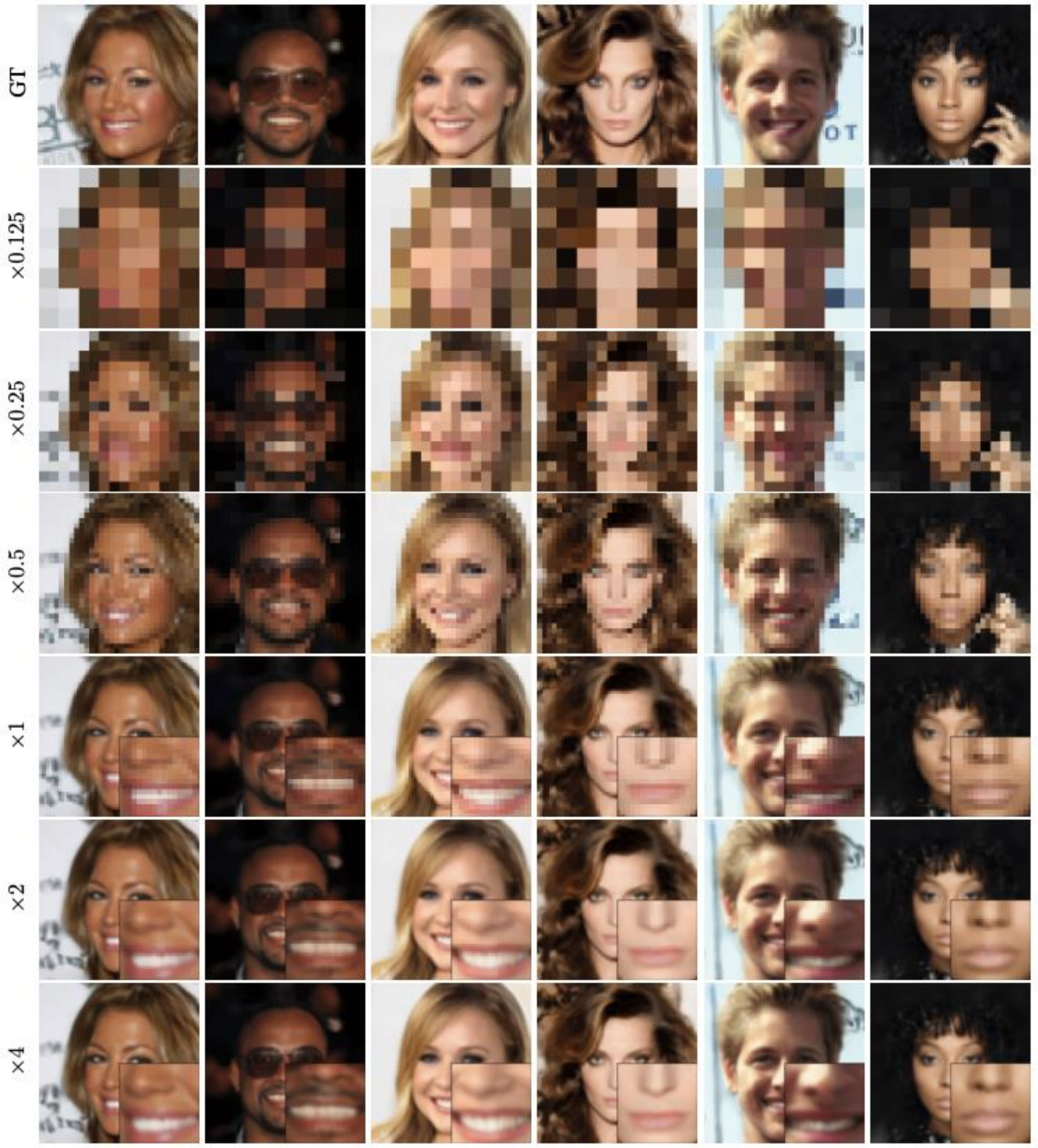
Model	PSNR (dB) ↑	FID ↓	HN Params ↓
CelebA-HQ (64×64)			
GASP [Dupont et al., 2022a]	-	**7.42**	25.7M
Functa [Dupont et al., 2022b]	$\leq$ **30.7**	40.40	-
VAMoH [Koyuncu et al., 2023]	23.17	66.27	25.7M
LDMI	24.80	18.06	**8.06M**
ImageNet (256×256)			
Spatial Functa [Bauer et al., 2023]	$\leq$ 38.4	$\leq$ 8.5	-
LDMI	20.69	**6.94**	102.78M

Table 1: Metrics on CelebA-HQ and ImageNet.

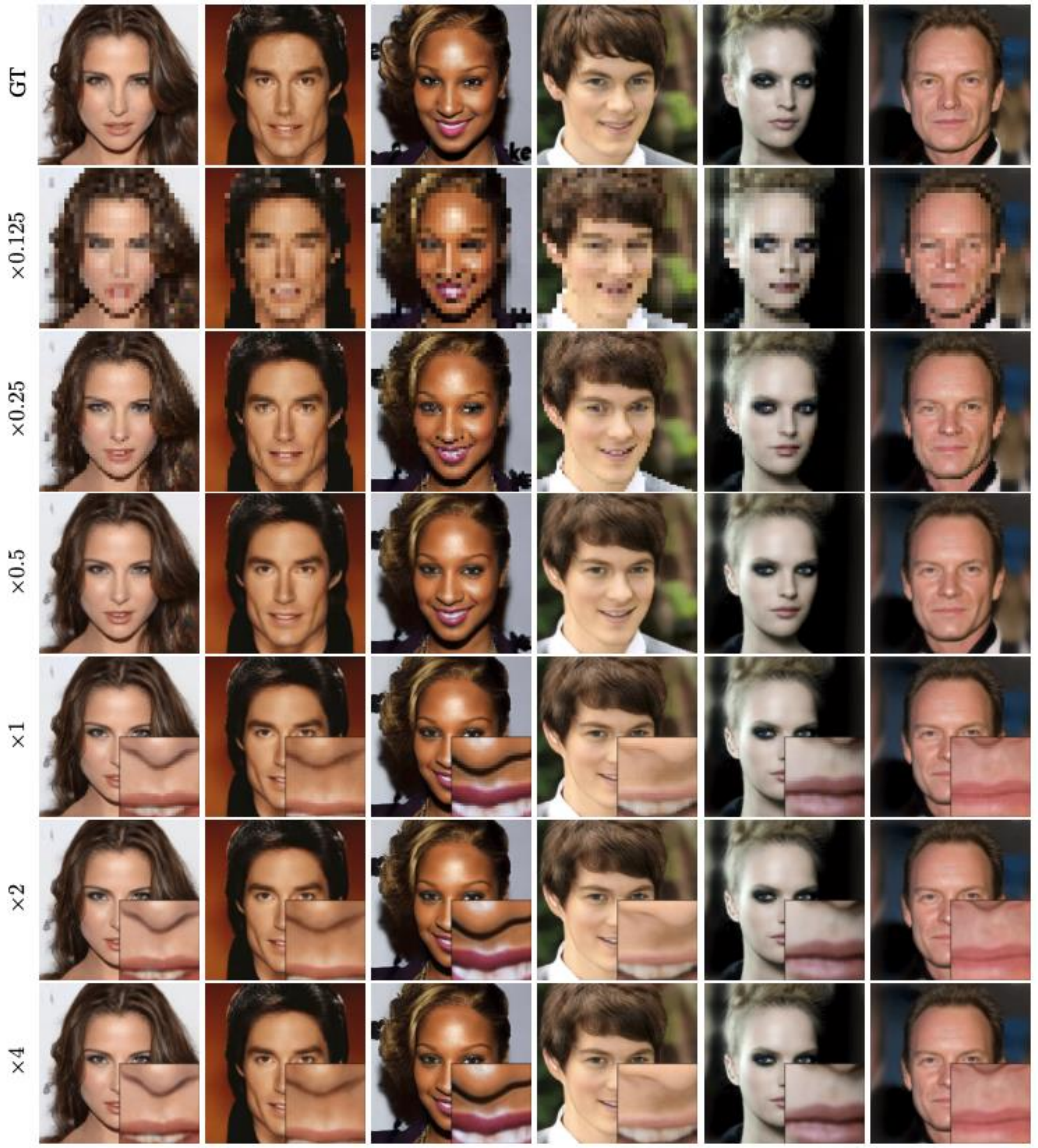
Experiments

Reconstruction

CelebA-HQ (64x64)



CelebA-HQ (256x256)



Model	Chairs (PSNR) ↑	ERA5 (PSNR) ↑
Functa [Dupont et al., 2022b]	29.2	34.9
VAMoH [Koyuncu et al., 2023]	38.4	39.0
LDMI	**38.8**	**44.6**

Table 2: Reconstruction quality (PSNR in dB) on ShapeNet Chairs and ERA5 climate data, demonstrating LDMI's strong generalization capabilities across modalities. Note that GASP is omitted as it is not applicable to INR reconstruction tasks.

Experiments

Data completion

VAMoH

LDMI

Experiments

Parameter efficiency

Method	HN Params	INR Weights	Ratio (INR/HN)
GASP/VAMoH	25.7M	50K	0.0019
LDMI	**8.06M**	**330K**	**0.0409**

Table 3: Parameter efficiency of hypernetworks (HN) in GASP/VaMoH and LDMI.

Method	HN Params	PSNR (dB)
LDMI-MLP	17.53M	24.93
LDMI-HD	**8.06M**	**27.72**

Table 4: Ablation study comparing MLP and hyper-transformer HD decoders on CelebA-HQ.

Conclusion

Thanks to learning **distributions of functions**, our proposed **VAMoH** can easily perform:

- Generation.
- Reconstruction.
- Conditional generation.
- Super resolution (interpolation).

While being:

✓ Robust to partially observed data.

✓ Expressive for generating high-quality data.

✓ Efficient in terms of inference.

Conclusion

Thanks to using **Latent Diffusion** and a **Transformer-based hypernetwork**, **LDMI** enhances

- Generation quality.
- Reconstruction accuracy.
- Conditional generation.
- Super resolution.

While:

✓ Being scalable.

✓ Being parameter efficient.

✓ Allowing for generation of **bigger INRs** and more complex data.

Further details

VARIATIONAL MIXTURE OF HYPERGENERATORS FOR LEARNING DISTRIBUTIONS OVER FUNCTIONS

Batuhan Koyuncu*
Saarland University
Saarbrücken, Germany

Pablo Sánchez-Martín
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Ignacio Peis
Universidad Carlos III de Madrid
Madrid, Spain

Pablo M. Olmos
Universidad Carlos III de Madrid
Madrid, Spain

Isabel Valera
Saarland University
Saarbrücken, Germany

[Paper]

[25] Koyuncu et at., 2023

Further details

HYPER-TRANSFORMING LATENT DIFFUSION MODELS

Ignacio Peis*
Technical University of Denmark

Batuhan Koyuncu
Saarland University

Isabel Valera
Saarland University

Jes Frellsen
Technical University of Denmark

[Paper]

[27] Peis et at., 2025

References

- [1] Campbell, A., Chen, W., Stimper, V., Hernandez-Lobato, J. M., & Zhang, Y. (2021, July). A gradient based strategy for hamiltonian monte carlo hyperparameter optimization. In *International Conference on Machine Learning* (pp. 1238-1248). PMLR.
- [2] Caterini, A. L., Doucet, A., & Sejdinovic, D. (2018). Hamiltonian variational auto-encoder. *Advances in Neural Information Processing Systems*, *31*.
- [3] Salimans, T., Kingma, D., & Welling, M. (2015, June). Markov chain monte carlo and variational inference: Bridging the gap. In *International conference on machine learning* (pp. 1218-1226). PMLR.
- [4] Ruiz, F. J., Titsias, M. K., Cemgil, T., & Doucet, A. (2021, December). Unbiased gradient estimation for variational auto-encoders using coupled Markov chains. In *Uncertainty in Artificial Intelligence* (pp. 707-717). PMLR.
- [5] Dupont, E., Whye Teh, Y. & Doucet, A.. (2022). Generative Models as Distributions of Functions. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research* 151:2989-3015.
- [6] Dupont, E., Kim, H., Eslami, S. A., Rezende, D. J., & Rosenbaum, D. (2022, June). From data to functa: Your data point is a function and you can treat it like one. In *International Conference on Machine Learning* (pp. 5694-5725). PMLR.

References

- [7] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- [8] Cremer, C., Li, X., & Duvenaud, D. (2018, July). Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning* (pp. 1078-1086). PMLR.
- [9] Bernardo, J. M. (1979). Expected information as expected utility. *the Annals of Statistics*, 686-690.
- [10] Ma, C., Tschiatschek, S., Palla, K., Hernández-Lobato, J. M., Nowozin, S., & Zhang, C. (2018). Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*.
- [11] Ma, C., Tschiatschek, S., Turner, R., Hernández-Lobato, J. M., & Zhang, C. (2020). VAEM: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, *33*, 11237-11247.
- [12] Child, R. (2020). Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*.

References

- [13] Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, *107*, 107501.
- [14] Mattei, P. A., & Frellsen, J. (2019, May). MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning* (pp. 4413-4423). PMLR.
- [15] Peis, I., Ma, C., & Hernández-Lobato, J. M. (2022). Missing Data Imputation and Acquisition with Deep Hierarchical Models and Hamiltonian Monte Carlo. arXiv preprint arXiv:2202.04599.
- [16] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, *69*(6), 066138.
- [17] Gong, W., Li, Y., & Hernández-Lobato, J. M. (2020). Sliced kernelized Stein discrepancy. *arXiv preprint arXiv:2006.16531*.
- [18] Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

References

- [19] Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, *79*(30), 2-4.
- [20] Sitzmann, V., Martel, J., Bergman, A., Lindell, D., & Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, *33*, 7462-7473.
- [21] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A. (2019). Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4460-4470).
- [22] Sitzmann, V., Zollhöfer, M., & Wetzstein, G. (2019). Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, *32*.
- [23] Ha, D., Dai, A. M., & Le, Q. V. HyperNetworks. In International Conference on Learning Representations.
- [24] Wu, W., Qi, Z., & Fuxin, L. (2019). Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 9621-9630).
- [25] Koyuncu, B., Sanchez-Martin, P., Peis, I., Olmos, P. M., & Valera, I. (2023). Variational Mixture of HyperGenerators for Learning Distributions Over Functions. In Proceedings of the 40th International Conference on Machine Learning, 2023.

References

- [26] Bauer, M., Dupont, E., Brock, A., Rosenbaum, D., Schwarz, J. R., & Kim, H. (2023). Spatial functa: Scaling functa to imagenet classification and generation. arXiv preprint arXiv:2302.03130.
- [27]: Peis, I., Koyuncu, B., Valera, I. & Frellsen, J. (2025). Hyper-Transforming Latent Diffusion Models. arXiv preprint arXiv:2504.16580.
- [28]: Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).
- [29]: Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In International Conference on Learning Representations.
- [30]: Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.
- [31]: Song, J., Meng, C., & Ermon, S. Denoising Diffusion Implicit Models. In International Conference on Learning Representations.

References

- [32]: Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations.

Thank you!

ipeaz@dtu.dk