Hyper-Transforming Latent Diffusion Models

Ignacio Peis, Batuhan Koyuncu, Isabel Valera, Jes Frellsen

DTU — Pioneer Centre for Artificial Intelligence — DDSA Danish Data Science Academy — Universität des Saarlandes — Zuse School ELIZA

Contributions

1. Novel Hyper-Transformer Decoder (HD) - first full Transformer-based probabilistic decoder for INR parameter generation.
2. Integration into Latent Diffusion Models with support for both full training and **hyper-transforming** paradigms.
3. Latent Diffusion Models for INRs (LDMI) offers scalable framework overcoming MLP-based hypernetwork bottlenecks while increasing expressiveness INRs.

Motivation

Challenges: Existing generative frameworks rely on structured representations that constrain resolution and generalization. MLP-based hypernetworks suffer from scalability bottlenecks when generating high-dimensional INRs, limiting flexibility and expressiveness for complex data.

Solution: LDMI combines Transformer-based hypernetworks with latent diffusion models for scalable, probabilistic INR generation.

Preliminaries

Implicit Neural Representations (INRs): Neural networks representing continuous functions with parameters Φ:

$$f_\Phi(\boldsymbol{x}) = \boldsymbol{y}, \quad \boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{y} \in \mathbb{R}^c \tag{1}$$

Hypernetworks: Networks generating parameters for other networks:

$$\Phi = g_\phi(\boldsymbol{z}) \mapsto f_\Phi(\boldsymbol{x}) = \hat{\boldsymbol{y}} \tag{2}$$

Latent Diffusion Models: Generative models applying diffusion in compressed latent space:

$$\mathcal{L}_{DDPM} = \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2\right] \tag{3}$$

Fast sampling via DDMI:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \begin{cases} \mathcal{N}\left(f_\theta^{(1)}(\boldsymbol{z}_1), \sigma_1^2 \boldsymbol{I}\right) & \text{if } t = 1 \\ q_\sigma\left(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, f_\theta^{(t)}(\boldsymbol{z}_t)\right) & \text{otherwise,} \end{cases} \tag{4}$$

LDMI

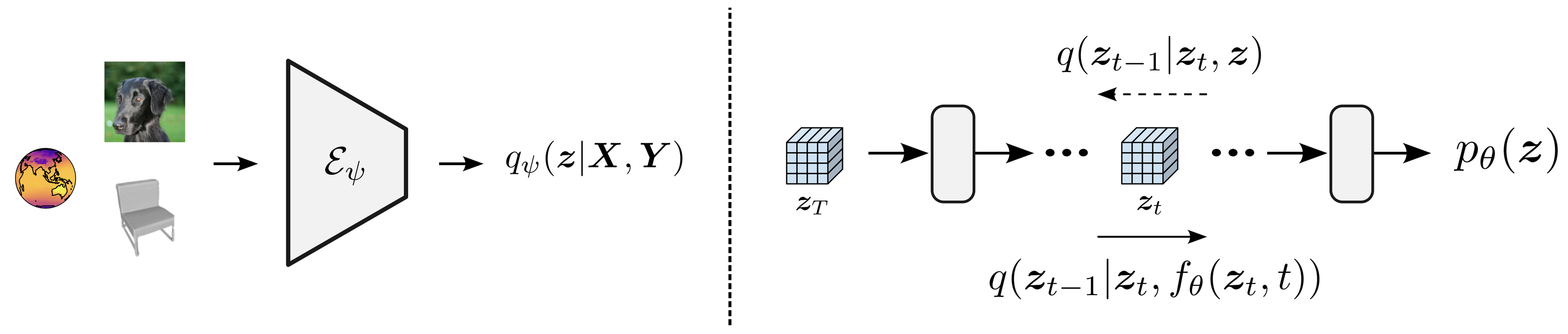


Figure 2: LDMI Encoder (left) and Latent Diffusion trained on latent space (right).

Approach: LDMI combines Transformer-based hypernetworks with latent diffusion models for scalable, probabilistic INR generation.

- **Encoder**: maps data to variational parameters $\boldsymbol{z}$.
- **Decoder**: full Transformer generates INR parameters via cross-attention and weight reconstruction.

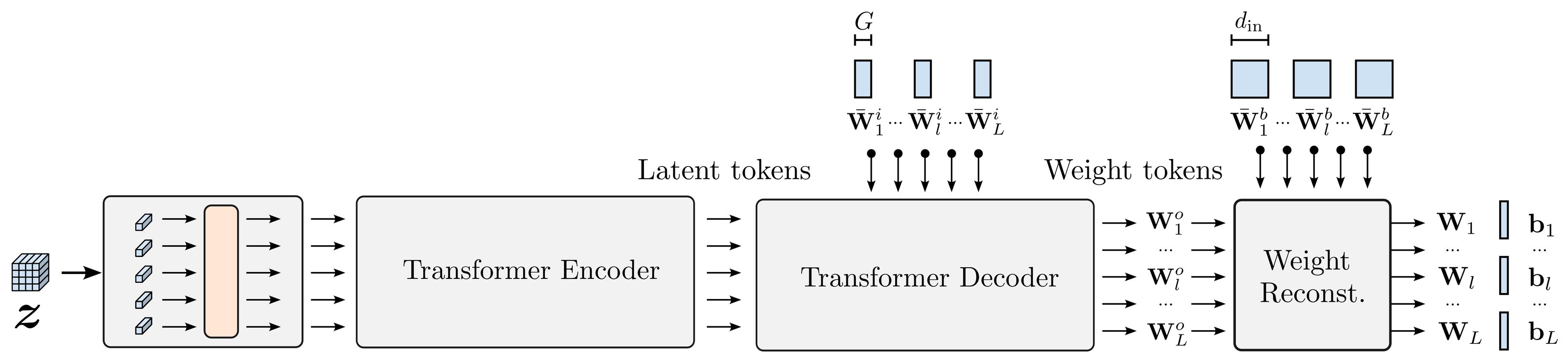


Figure 3: The HD Decoder

$$\mathcal{R}^{(\text{scale})}\left(\boldsymbol{w}^{\text{o}}_{[c/k]}, \bar{\boldsymbol{w}}^{\text{b}}_c\right) = \left(1 + \boldsymbol{w}^{\text{o}}_{[c/k]}\right) \odot \bar{\boldsymbol{w}}^{\text{b}}_c. \tag{5}$$

Training Paradigms:

1. Full Training:

First stage:

$$\mathcal{L}_{\text{VAE}}(\phi,\psi) = \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})}\left[\log p_\Phi(\boldsymbol{Y}|\boldsymbol{X})\right] - \beta \cdot D_{KL}(q_\psi(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y}) \| p(\boldsymbol{z})) \tag{6}$$

Second stage:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)\|^2\right] \tag{7}$$

2. Hyper-Transforming: Adapt pre-trained LDM by freezing $\{\psi, \theta\}$ and training decoder:

$$\mathcal{L}_{\text{HT}}(\phi) = \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})}\left[\log p_\Phi(\boldsymbol{Y}|\boldsymbol{X})\right], \tag{8}$$

Hyper-transforming *enables efficient adaptation of existing diffusion models without full retraining, leveraging pre-trained latent spaces.*

Results

Generation

LDMI achieves high-quality unconditional and conditional generation across multiple modalities and arbitrary resolutions.

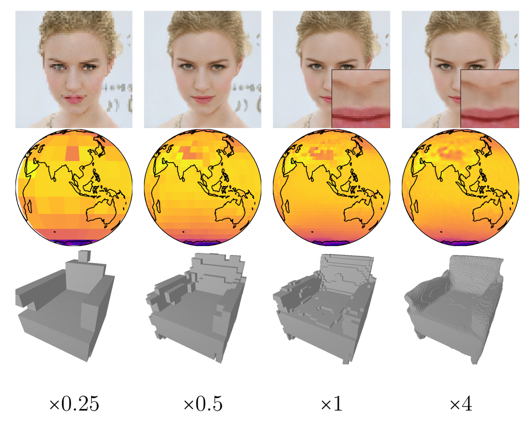


Figure 4: LDMI samples at multiple resolutions and modalities.

Figure 5: CelebA-HQ (64×64) samples from baselines and LDMI.

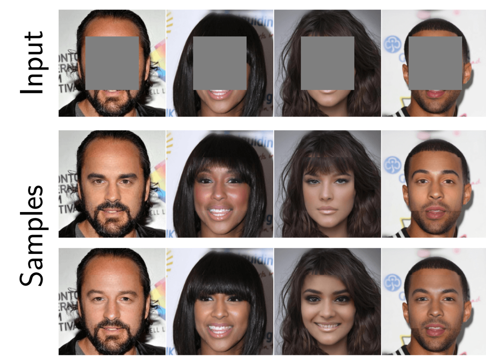
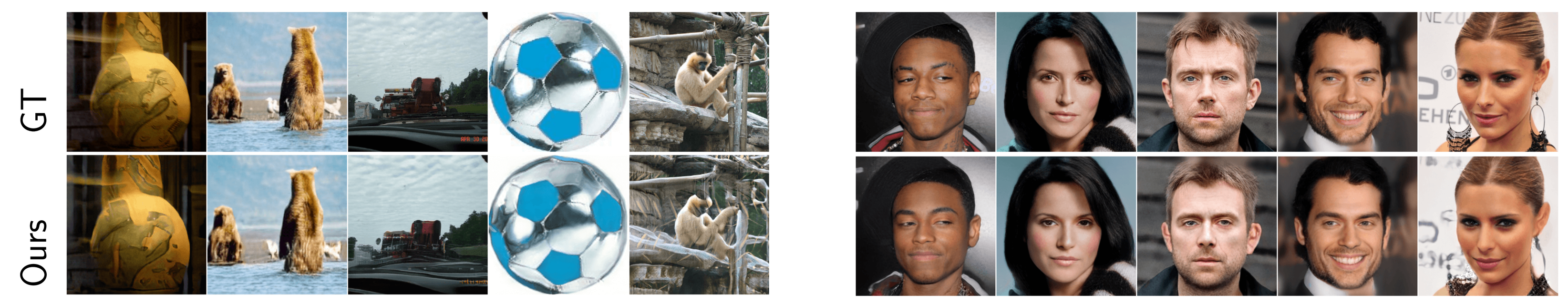


Figure 6: Inpainting with LDMI on CelebA-HQ (256×256).

Reconstruction:

Our framework models the space of INRs to represent data at significantly higher resolutions, beyond the capabilities of existing methods.



(a) ImageNet (b) CelebA-HQ (256×256)

Figure 7: Reconstructions by our LDMI trained by hyper-transforming pre-trained LDMs.

Quantitative Results:

We generate INRs that are approximately $7\times$ larger in size using hypernetworks with less than $1/3$ the number of parameters.

Model	PSNR (dB) ↑	FID ↓	HN Params ↓
CelebA-HQ (64×64)			
GASP	-	**7.42**	25.7M
Functa	≤ **30.7**	40.40	-
VAMoH	23.17	66.27	25.7M
LDMI	24.80	18.06	**8.06M**
ImageNet (256×256)			
Spatial Functa	≤ 38.4	≤ 8.5	-
LDMI	20.69	**6.94**	102.78M

Table 1: Metrics on CelebA-HQ and ImageNet.

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

Table 2: Scalability analysis.

Model	Chairs (acc %) ↑	ERA5 (PSNR dB) ↑
Functa	99.51	34.9
VAMoH	96.75	39.0
LDMI	97.25	44.6

Table 3: Reconstruction quality on ShapeNet Chairs and ERA5.