Tokenized Neural Fields: Structured Representations of Continuous Signals

Azmi Haider
Department of Computer Science
University of Haifa
Haifa, Israel
ahaide03@campus.haifa.ac.il

Dan Rosenbaum
Department of Computer Science
University of Haifa
Haifa, Israel
danro@cs.haifa.ac.il

Abstract

We introduce Tokenized Neural Fields (TNF), a unified framework for representing continuous signals through a compact set of learnable tokens. Unlike encoder-based pipelines or global latent codes, TNF provides a structured tokenization in which individual tokens specialize to distinct aspects of a signal and interact with coordinate queries via cross-attention. This decoupling of representation from decoder architecture enables scalable training across modalities, efficient adaptation to new signals, and a natural basis for probabilistic inference in token space. We validate TNF across 1D function regression, 2D image reconstruction, and 3D scene modeling, showing that tokenized representations achieve superior fidelity with fewer parameters compared to encoder- or latent-based baselines. Beyond accurate reconstructions, TNF tokens exhibit emergent specialization, support interpolation and morphing, and enable generative modeling when paired with diffusion transformers. Together, these results highlight tokenization as a powerful paradigm for bridging implicit neural representations with the structured inference and generative capabilities increasingly central to large foundation models.

1 Introduction

Learning compact and generalizable representations of continuous signals—such as functions, images, and 3D scenes—remains a central challenge in machine learning and generative modeling. Classical approaches rely on modality-specific encoders that map data into latent spaces for reconstruction or generation. While effective, these pipelines are tightly coupled to architectural choices and computationally expensive to scale.

Implicit Neural Representations (INRs), or Neural Fields (NFs), offer an alternative by modeling signals as continuous functions from coordinates to values. Examples include SIREN [Sitzmann et al., 2020a] for high-frequency image reconstruction and NeRF [Mildenhall et al., 2020] for 3D scene rendering. However, most INRs require per-instance training, limiting scalability. Conditional variants such as Functa [Dupont et al., 2022] improve efficiency by sharing a decoder across instances, but compress each signal into a *global latent*, which can discard localized structure and harm performance [Bauer et al., 2023].

Modern foundation models instead rely on *tokenized representations*: sets of tokens that capture structure and enable inference, reasoning, and generation. While tokenization arises naturally in discrete domains like language, adapting it to continuous signals has typically required handcrafted discretizations (e.g., image patches or voxel grids). This motivates our central question: *can continuous signals be tokenized into compact, learnable sets of tokens that provide structured representations for reconstruction, generation, and probabilistic inference?*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Structured Probabilistic Inference & Generative Modeling..

We propose **Tokenized Neural Fields (TNF)**, a unified framework that represents each signal with a small, learnable token set and a shared coordinate-conditioned decoder. Tokens are optimized per instance to capture structure, while the decoder is shared across all instances. This decoupling yields scalable training, efficient adaptation, and structured token spaces that support both high-quality reconstruction and generative modeling.

We validate TNF on:

- **1D functions**, where tokens recover smooth Gaussian Process samples;
- **2D images**, where tokens yield sharp reconstructions and support generative modeling with diffusion transformers, outperforming encoder-based baselines with fewer parameters;
- **3D scenes**, where tokenized representations enable geometry-free, consistent novel view synthesis and generative modeling with diffusion transformers.

Together, these results suggest that tokenization offers a compact, expressive, and scalable alternative to conventional INRs, while opening new directions for probabilistic inference and generative modeling in the era of foundation models.

2 Related Work

Neural Fields (NFs) have emerged as a powerful paradigm for modeling continuous signals such as images, functions, and 3D scenes. These methods learn a mapping from input coordinates to output values using neural networks, often achieving high-fidelity reconstruction and compact signal encoding. However, many existing NF approaches rely on per-instance optimization, structured encoders, or modality-specific pipelines, limiting their scalability, generalization, and reuse in downstream tasks.

Probabilistic models of continuous signals. Early work on modeling continuous signals with neural networks focused on probabilistic formulations inspired by Gaussian Processes (GPs). Conditional Neural Processes (CNPs) [Garnelo et al., 2018a] and Neural Processes (NPs) [Garnelo et al., 2018b] aim to learn distributions over functions conditioned on context points, blending ideas from GPs and deep learning. Extensions such as attentive-NP [Kim et al., 2019] and Transformer-NP [Nguyen and Grover, 2022] introduce attention mechanisms for improved uncertainty modeling and generalization. These approaches operate directly on sets of (x, y) pairs and emphasize probabilistic inference, whereas our goal is to produce *structured, tokenized summaries* that are reusable across tasks and modalities while still supporting probabilistic reasoning.

Per-instance NF training. A classical approach is to train a separate NF per signal. For instance, SIREN [Sitzmann et al., 2020a] represents high-frequency signals with sinusoidal activations, while NeRF [Mildenhall et al., 2020] learns scene-specific radiance fields via volumetric rendering. These methods achieve impressive fidelity but scale poorly, as each new signal requires training a new network. In contrast, our method employs a shared decoder with per-instance tokens, enabling generalization without retraining full models.

Learned initialization for NFs. Meta-learning strategies such as Learned Init [Tancik et al., 2020] accelerate per-instance optimization by providing better initializations. While this reduces training time, it does not yield compact, transferable representations—highlighting the need for more structured and reusable parameterizations.

Conditional NFs and auto-decoding. Shared decoders conditioned on latent vectors have been explored in works like Functa [Dupont et al., 2022] for general signals and Gaudi [Bautista et al., 2022], Single-stage Diffusion NeRF [Chen et al., 2023], and Diffusion Prior for NeRFs [Yang et al., 2023] for 3D scenes. These models follow the *auto-decoding* paradigm [Bojanowski et al., 2019, Park et al., 2019], where latents are optimized jointly with a shared decoder. However, these latents are typically global and structure-less. Spatial-Functa [Bauer et al., 2023] highlighted this limitation by proposing spatially structured representations. Our work takes this further by showing that *learned tokenization*—rather than imposing fixed structure—provides an emergent organization of the representation space, where tokens specialize to different parts of a signal.

Attention-based models. Transformers have been applied to INRs in several forms. TransINR [Chen and Wang, 2022] uses meta-learning to generalize across tasks, IPC [Kim et al., 2023] introduces

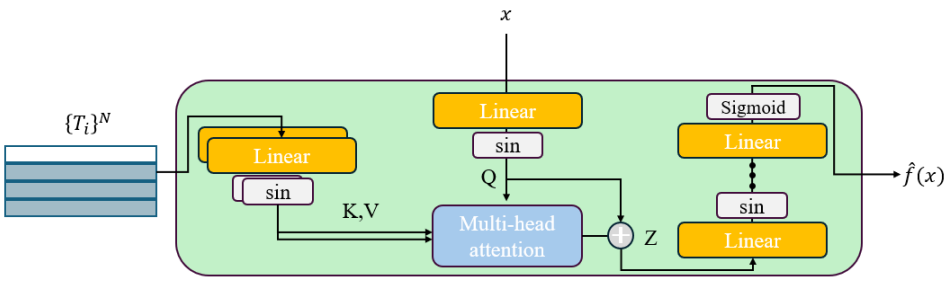

Figure 1: TNF: Our token-based coordinate-attention decoder. Each function f is represented by a different set of tokens $\{T_i\}^N$, and can be evaluated at point x using the decoder.

compositional instance pattern encoding to encourage reusability, and ANR [Zhang et al., 2024] leverages attention to generate R-tokens through a heavy encoder-decoder hypernetwork. In 3D, several recent works explore token-based scene representations, including OSRT [Sajjadi et al., 2022], 3DShape2VecSet [Zhang et al., 2023], and equivariant neural fields [Wessels et al., 2025], which rely on explicit encoders to map raw data into tokens. These approaches demonstrate the promise of attention and tokenization for structured correspondence, but still depend on handcrafted encoders or large auxiliary networks. In contrast, our formulation directly *learns tokens via gradient-based optimization*, yielding compact, reusable, and structured representations with far fewer parameters, and without requiring handcrafted design.

3 Tokenized Neural Fields (TNF)

We propose a unified framework for learning compact and flexible signal representations using a small set of learnable tokens and a shared coordinate-conditioned decoder. Each instance—whether a function, image, or 3D scene—is represented by a fixed set of latent tokens. These tokens are optimized to reconstruct the underlying signal via a shared decoder network, enabling instance-specific adaptation while maintaining global generalization. This formulation removes the need for modality-specific encoders or per-instance networks and supports scalable, transferable representations across domains.

3.1 Overview

Let $f : \mathbb{R}^d \to \mathbb{R}^c$ denote a continuous signal (e.g., a 1D curve, a 2D image, or a 3D scene). Rather than fitting a dedicated neural network for each f, we associate it with a learnable token set:

$$T = \{T_n\}_{n=1}^N, \quad T_n \in \mathbb{R}^{d_t}$$

A shared decoder D_θ maps coordinates and tokens to signal outputs:

$$\hat{f}(x) = D_\theta(x, T)$$

The decoder is trained across a dataset of signals, while the token sets are optimized per instance. During inference, the decoder remains fixed and only the tokens are adapted for new data—enabling efficient specialization with minimal compute.

3.2 Decoder Architectures

Figure. 1 illustrates our decoder architecture. The decoder maps a coordinate $x \in \mathbb{R}^{d_x}$ and the token set T to the signal prediction $\hat{f}(x)$. Its design is flexible and can adapt to various data modalities. A natural choice when using token-based representation is to use cross-attention between the coordinate x and the signal's tokens. The representation tokens T_i are concatenated with sequential sinusoidal positional encoding and together with the query coordinate x undergo a linear transformation followed by a sine activation function. Then, cross attention is performed between the coordinate and the tokens, followed by several linear layers coupled with sine activations, except for a sigmoid activation in the last layer. We follow the standard transformer and use layer normalization on the tokens before the cross attention, a residual connection between the query and the attention output and layer normalization after the residual connection. We use this approach throughout our experiments, where

our decoders mainly differ in the number of layers, their dimensions and the dimensions of the input x and output $\hat{f}(x)$.

1D signals. For scalar functions $f : \mathbb{R} \to \mathbb{R}$, the coordinate x is used as a scalar query in a cross-attention module over the token set. The resulting embedding is then decoded through several layers , as depicted in Figure. 1 ,and projected back to a 1D scalar output. Essentially, the decoder is a mapping between a sequence representation to a continuous signal.

2D signals: images. Images are examples of complex continuous signals over 2D spaces, mapping pixel coordinates to color. In our decoder, each coordinate $x = (u, v) \in [0, 1]^2$ is used as a query into a cross-attention layer over the image-specific token set. The resulting attended feature is decoded by an MLP to predict RGB values. This structure allows the model to generalize across images while preserving spatial precision via coordinate querying and instance adaptation via tokens.

High dimensional signals: light fields. Light fields [Sitzmann et al., 2021] is a method to represent multiple views of 3D scenes. In this representation, pixels from the different views are represented by a ray $x = (\mathbf{o}, \mathbf{d}) \in \mathbb{R}^6$ (camera origin and direction). Essentially this is a continuous signal over a 6-dimensional space, mapping the ray coordinate to a 3D color value. In our decoder, this ray representation is used to query the token set via cross-attention. To reconstruct an image of a 3D scene from a certain view, we query the model with the light field representation of all the pixels in the image. For each query, the output of the cross-attention is decoded into a pixel color. This enables reconstructing images of arbitrary views of a 3D scene without using geometry, encoders, or 3D grids. This method aligns closely with the 1D/2D cases, showing that TNF supports fully flexible geometry-free rendering from tokens.

3.3 Cross-Attention Querying

Cross-attention enables dynamically extracting relevant content from the token set into different locations in the continuous signal. Given a query coordinate $x \in \mathbb{R}^{d_x}$, we compute:

$$
\begin{aligned}
Q &= \phi_q(x), \quad Q \in \mathbb{R}^{1 \times d_h} \\
K &= \phi_k(T), \quad K \in \mathbb{R}^{N \times d_h} \\
V &= \phi_v(T), \quad V \in \mathbb{R}^{N \times d_h}
\end{aligned}
$$

Here, $T \in \mathbb{R}^{N \times d_t}$ is the set of learnable tokens for an instance, d_t is the token dimensionality, d_h is the attention hidden dimension, and N is the number of tokens.

We then compute the scaled dot-product attention:

$$
\begin{aligned}
A &= \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right), \quad A \in \mathbb{R}^{1 \times N} \\
Z &= AV + Q, \quad Z \in \mathbb{R}^{1 \times d_h}
\end{aligned}
$$

The attention output Z is passed through a small MLP to produce the final prediction (e.g., RGB value or scalar signal). This querying formulation follows the original Transformer attention mechanism Vaswani et al. [2017], adapted to continuous-coordinate domains through instance-specific token sets.

3.4 Training and Adaptation

Training is done end-to-end across a dataset of M signals following the *auto-decoding* approach Bojanowski et al. [2019], Park et al. [2019]. For each instance f_i, the token set $T^i = \{T^i_n\}_{n=1}^N$ is randomly initialized and jointly optimized with the shared decoder D_θ using:

$$
\min_{\theta,\, \{T^i\}_{i=1}^M} \frac{1}{M} \sum_{i=1}^{M} \frac{1}{K} \sum_{j=1}^{K} \left\| f_i(x_j) - D_\theta(x_j, T^i) \right\|^2
$$

where $\{x_j\}_{j=1}^K$ are randomly sampled coordinates.

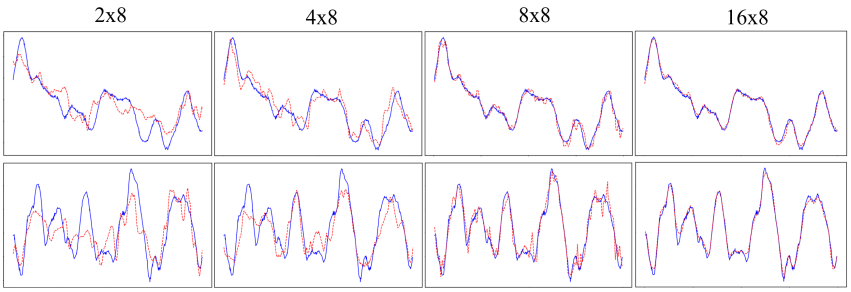

Figure 2: 1D function reconstruction using TNF. Ground truth functions (blue) are sampled from Gaussian Processes with RBF kernel (top row) and Matérn kernel (bottom row). TNF reconstructions are shown in red. Columns correspond to representation with increasing token counts (2×8, 4×8, 8×8, and 16×8), demonstrating improved fidelity with more tokens.

A core design principle of TNF is to separate instance-specific and instance-agnostic components: the coordinate input x_j is shared across all instances, while the token set T^i is unique to each instance. Cross-attention bridges these two spaces, allowing the model to specialize per instance while sharing a common decoder.

At inference time, the decoder is kept fixed and new tokens are adapted to unseen data using gradient-based optimization.

4 Experiments

We evaluate our Token-conditional Neural Fields (TNF) across 1D function regression, 2D image modeling, and 3D scene reconstruction. Our goal is to demonstrate the generality, compactness, and flexibility of token-based representations using a shared decoder and per-instance tokens.

4.1 1D Functions

We begin by modeling 1D functions sampled from Gaussian Processes using RBF and Matérn kernels. 1K functions are used for training. Each function $f : [0, 1] \to \mathbb{R}$ is represented using a compact set of learnable tokens ($N \times d_t$), trained for 150 steps over 400 sampled points. We use a decoder with 8-head cross attention, hidden dimension of 64 and and 5 output layers.

Figure 2 shows reconstructions with increasing token counts (2×8, 4×8, 8×8, and 16×8), illustrating that higher token capacity yields more accurate representations of the target signal.

4.2 2D Image Modeling

We evaluate TNF on 2D image reconstruction using the CelebA [Liu et al., 2015] and ImageNette [Howard, 2019] datasets, and compare it to recent encoder-based baselines including TransINR, IPC, ANR, and Learned Init. In contrast to these methods, TNF directly optimizes per-instance tokens and uses a shared decoder—resulting in significantly fewer parameters without sacrificing visual fidelity. We use the same decoder architecture as for 1D with dimensions 256 for the attention, and 512 for the output layers. For both CelebA and Imagenette we train on 10K images of 178×178 and evaluate on 100 random held out images.

As shown in Figure 3, TNF achieves higher PSNR scores than all baselines, with up to 4–10$\times$ fewer representation parameters. Notably, the reconstructions produced by TNF are not only quantitatively superior but also exhibit **high sharpness and perceptual quality**. Even with compact token sets, the model captures fine details such as facial contours, texture, and edges—outperforming deeper and heavier encoder-based pipelines.

Token specialization and generative modeling. TNF tokens exhibit both structured specialization and generative potential. Figure 4 shows cross-attention heatmaps from a CelebA model trained with 16 tokens: despite no supervision, tokens consistently attend to coherent regions (e.g., hair, background, face), indicating emergent structure. Figure 5 shows image samples generated by

Dataset	Method	PSNR	$N \times d_t$	Rep (K)	Model (M)
CelebA	Learned Init	30.37	–	199	0.2
	TransINR	33.33	256×259	66	43
	IPC	35.96	256×256	66	43
	ANR	35.91	512×256	131	54
	TNF	31.5	64×32	2	2
	TNF	**37.02**	512×16	**8**	2
	TNF	**42.14**	512×32	**16**	2
ImageNette	Learned Init	27.07	–	199	0.2
	TransINR	29.77	256×259	66	43
	IPC	38.46	256×256	66	43
	ANR	40.30	512×256	131	54
	TNF	**44.40**	1024×32	**32**	3
	TNF	**47.98**	1024×64	**66**	7

Figure 3: Comparison of reconstruction performance. (Left) Quantitative results on CelebA and ImageNette. TNF achieves higher PSNR with much smaller representation and model size compared to prior methods. (Right) Qualitative reconstructions with 512×32 tokens: Left column: CelebA; right column: ImageNette. Image pairs show ground truth (left) vs. TNF output (right).

Figure 4: Token specialization: cross-attention heatmaps show consistent, spatially coherent focus for individual tokens across CelebA images.

Figure 5: Generative modeling: a DiT trained on compact token sets produces diverse, coherent images decoded by TNF.

training a Diffusion Transformer (DiT) [Peebles and Xie, 2022] directly over compact token sets (64×32). The generated tokens, decoded by our shared network, yield diverse and coherent images, demonstrating that even lightweight tokenizations support effective generative modeling.

4.3 3D Scene Reconstruction

We evaluate TNF on neural scene reconstruction using the car category from the ShapeNet SRN dataset [Sitzmann et al., 2020b]. Each scene is represented by a compact set of learned tokens, decoded via the light field representation decoder.

The dataset provides RGB images from multiple viewpoints around each object. During training, we randomly sample a batch of B rays per step, where each ray is defined by a camera origin and direction and is supervised with the corresponding ground-truth pixel color.

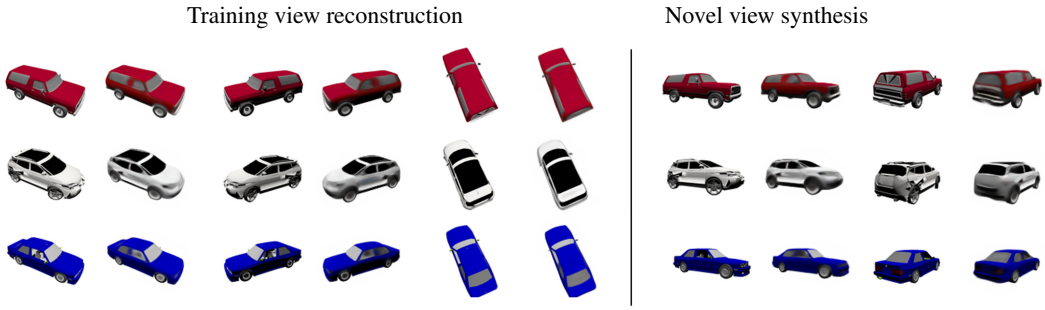

Figure 6: 3D view reconstruction and synthesis. Each pair shows the ground truth view (left) and the reconstructed view (right).

Figure 7: Linear interpolation in token space between two scenes. Shown are two examples (rows) from different viewpoints; in each, the scene on the left smoothly transitions through 10 interpolation steps into a different scene on the right, with gradual changes in geometry and appearance.

Each 3D scene is represented by a set of tokens of size 512×32. During training, each ray is defined by its camera origin $\mathbf{o}$ and direction $\mathbf{d}$. These inputs, along with the tokens, are independently encoded using three-layer MLPs with hidden dimension 512 and ReLU activations.

The encoded ray features are concatenated and used as queries in a stack of 8 cross-attention blocks. Each block consists of 8-head multi-head attention with hidden dimension 512, followed by two-layer feedforward networks with GELU activations. The attention mechanism computes interactions between the ray embeddings and the scene tokens, enabling token-conditioned reasoning over the ray space.

The output of the attention stack is then passed through a 7-layer MLP with ReLU activations, followed by a final sigmoid layer to predict the RGB color of the pixel corresponding to that ray.

Once optimized, the tokens serve as a compact and reusable representation of the 3D scene. They can be used to reconstruct training views or generate novel renderings from unseen camera poses. For more details, see Appendix A.2

Qualitative Results. Figure 6 shows both training view reconstruction and novel view synthesis results. Despite the compact token representations, the method produces coherent and detailed reconstructions. These outputs are generated without any 3D encoder or mesh supervision, with a shared decoder used across all scenes—demonstrating the strength and flexibility of TNF.

Token Interpolation and Generative Modeling. To analyze the expressiveness of the token space, we consider both interpolation and generation. Figure 7 shows linear interpolation between token sets from two scenes, producing smooth transitions in geometry and color, suggesting that the token space forms a continuous manifold suitable for editing and morphing. Figure 8 shows samples from a Denoising Diffusion Probabilistic Model (DDPM) [Ho et al., 2020] trained directly over compact scene tokens (64×32), decoded through our shared triplane decoder. The generated tokens yield coherent multi-view 3D scenes with diverse geometry and appearance, highlighting both the expressiveness and generative potential of TNF.

Token specialization. Figure 9 illustrates that even in 3D, individual tokens tend to focus on distinct semantic parts of the scene. Each row corresponds to a specific token, and columns show different car scenes from the SRN dataset. Notably, despite being trained without supervision, tokens specialize in modeling particular regions (e.g., roof, wheels, or headlights), supporting the emergence of structured and interpretable token-space behavior in volumetric domains.

Figure 8: 3D scene samples generated from a DDPM trained on TNF tokens producing diverse and coherent 3D structures.

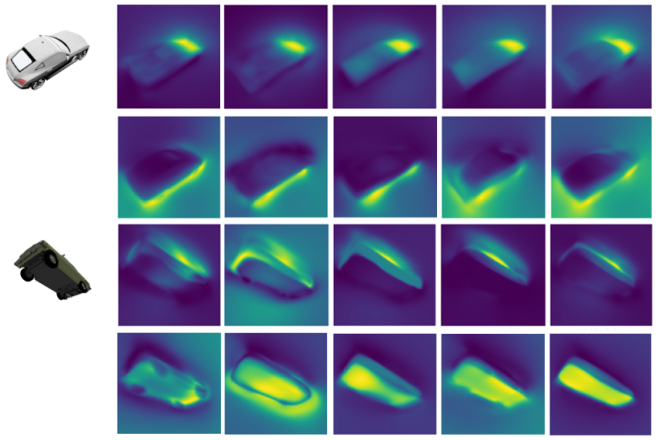

Figure 9: Visualization of token specialization in 3D. Each row corresponds to a different token, highlighting the regions of the car it attends to across multiple scenes (columns). The colored images on the left indicate the viewing angles of the corresponding scenes.

5 Conclusion

We introduced **Tokenized Neural Fields (TNF)**, a unified framework for representing continuous signals using compact sets of learnable tokens and a shared coordinate-conditioned decoder. Across 1D functions, 2D images, and 3D scenes, TNF achieves high-fidelity reconstruction with fewer parameters than encoder-based or global-latent approaches, while exhibiting emergent token specialization and supporting generative modeling.

Our experiments highlight three advantages: (i) *scalability*, by decoupling tokenized representations from the decoder architecture and enabling adaptation with minimal compute; (ii) *emergent structure*, as tokens consistently specialize to coherent regions or semantic parts without explicit supervision; and (iii) *compactness*, achieving strong reconstructions with significantly fewer learnable parameters. At the same time, TNF inherits some limitations: adaptation requires gradient-based optimization, which introduces latency compared to encoder-based one-shot inference, and the token space may benefit from stronger inductive biases or amortized inference strategies [Marino et al., 2018, Dupont et al., 2022].

Looking forward, we believe the greatest potential of TNF lies in *probabilistic inference over token spaces*. Reconstruction can be reframed as posterior inference, where the goal is to recover plausible token sets from incomplete or uncertain observations. Learning priors over tokens would enable posterior sampling, uncertainty quantification, and hallucination in under-constrained regimes. Rather than relying solely on deterministic optimization, amortized inference or conditional diffusion models [Ho et al., 2020] could yield uncertainty-aware scene completion, novel view synthesis, and probabilistic reasoning about 3D structure.

In this view, tokenization is not only a compact encoding strategy, but also a foundation for structured probabilistic inference and generative modeling in continuous domains—bridging implicit neural fields with the token-based reasoning paradigms of large foundation models.

References

Matthias Bauer, Emilien Dupont, Andy Brock, Dan Rosenbaum, Jonathan Richard Schwarz, and Hyunjik Kim. Spatial functa: Scaling functa to imagenet classification and generation, 2023.

Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Josh Susskind. Gaudi: A neural architect for immersive 3d scene generation, 2022.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the Latent Space of Generative Networks, May 2019. URL http://arxiv.org/abs/1707.05776.

Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction, 2023. URL https://arxiv.org/abs/2304.06714.

Yinbo Chen and Xiaolong Wang. Transformers as meta-learners for implicit neural representations, 2022. URL https://arxiv.org/abs/2208.02801.

Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo J. Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you should treat it like one. *CoRR*, abs/2201.12204, 2022. URL https://arxiv.org/abs/2201.12204.

Marta Garnelo, Dan Rosenbaum, Chris Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. *ICML*, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes, 2018b. URL https://arxiv.org/abs/1807.01622.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Jeremy Howard. Imagenette: A smaller subset of imagenet. https://github.com/fastai/imagenette, 2019. Accessed: 2025-08-25.

Chiheon Kim, Doyup Lee, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Generalizable implicit neural representations via instance pattern composers, 2023. URL https://arxiv.org/abs/2211.13223.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes, 2019. URL https://arxiv.org/abs/1901.05761.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3403–3412. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/marino18a.html.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16569–16594. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nguyen22b.html.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation, 2019. URL https://arxiv.org/abs/1901.05103.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Mehdi S. M. Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetić, Mario Lučić, Leonidas J. Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer, 2022. URL https://arxiv.org/abs/2206.06922.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *CoRR*, abs/2006.09661, 2020a. URL https://arxiv.org/abs/2006.09661.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations, 2020b. URL https://arxiv.org/abs/1906.01618.

Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Frédo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *CoRR*, abs/2106.02634, 2021. URL https://arxiv.org/abs/2106.02634.

Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. *CoRR*, abs/2012.02189, 2020. URL https://arxiv.org/abs/2012.02189.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

David R Wessels, David M Knigge, Samuele Papa, Riccardo Valperga, Sharvaree Vadgama, Efstratios Gavves, and Erik J Bekkers. Grounding continuous representations in geometry: Equivariant neural fields, 2025. URL https://arxiv.org/abs/2406.05753.

Guandao Yang, Abhijit Kundu, Leonidas J. Guibas, Jonathan T. Barron, and Ben Poole. Learning a diffusion prior for nerfs, 2023. URL https://arxiv.org/abs/2304.14473.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models, 2023. URL https://arxiv.org/abs/2301.11445.

Shuyi Zhang, Ke Liu, Jingjun Gu, Xiaoxu Cai, Zhihua Wang, Jiajun Bu, and Haishuai Wang. Attention beats linear for fast implicit neural representation generation, 2024. URL https://arxiv.org/abs/2407.15355.

A Appendix

A.1 2D image modeling

CelebA (10k images) and ImageNette (10 classes, each roughly has 1k images) datasets are center-cropped and resized to 178×178 RGB images.

The model is trained using two separate Adam optimizers: one for the decoder parameters with a learning rate of 8^{-4}, and another for the tokens with a learning rate of 2^{-5}. Both learning rates decay with a factor of 0.999.

DiT for 2D Token Generation. In our 2D generative modeling experiment, we train a DiT model [Peebles and Xie, 2022] with 28 transformer layers, a hidden size of 1152, and 16 attention heads. The model is optimized using the AdamW optimizer with a learning rate of 1^{-4}. DiT operates directly in the token space, learning to generate compact token representations which are then decoded into images using our shared coordinate-based decoder.

As shown in Figure 10, the generated samples (right) and their nearest reconstructions from the training set (left) exhibit clear visual differences, indicating that the model does not memorize but instead learns to produce diverse and novel outputs.

Figure 10: Generated token samples (right) and their nearest training-set reconstructions (left). The visual differences indicate the model generates novel and diverse samples rather than memorizing the training data.

Token specialization To complement the results shown in the main paper, we visualize the full set of token attention maps. Figure 11 displays the activation heatmaps of 16 tokens (rows) across 12 test scenes (columns), highlighting how each token consistently attends to specific regions. This reinforces the observation that token specialization emerges naturally, even without spatial supervision.

A.2 3D scene representation modeling

The ShapeNet SRN car dataset provides 3k scenes of cars, each has 250 images from different camera positions split evenly to train and test. The scenes contain RGB images of size 128×128. During training, we randomly sample a batch of $B = 4096$ rays from all training images per step. The model is trained using two separate Adam optimizers: one for the decoder parameters and one for the tokens, both with a learning rate of 1^{-3}, and decay with a factor of 0.999.

DiT for 3D Token Generation. For 3D generative modeling, we use the same DiT framework as in 2D but with 12 transformer layers, a hidden size of 768, and 12 attention heads. The model is trained directly in the token space to generate compact 3D scene tokens, which are decoded by our shared decoder. Figure 12 compares generated samples with their nearest neighbors in the training set, measured in token space. While neighbors share coarse similarities, clear geometric and appearance differences remain, indicating that the model synthesizes novel and diverse scenes rather than overfitting to training examples.

Figure 11: Full attention heatmaps for 16 tokens (rows) across 12 test scenes (columns). Each token consistently focuses on similar regions across different images, demonstrating spatial specialization consistency.

Figure 12: Generated token samples (right) and their nearest training-set neighbors (left). The visual differences indicate the model generates novel and diverse samples rather than memorizing the training data.