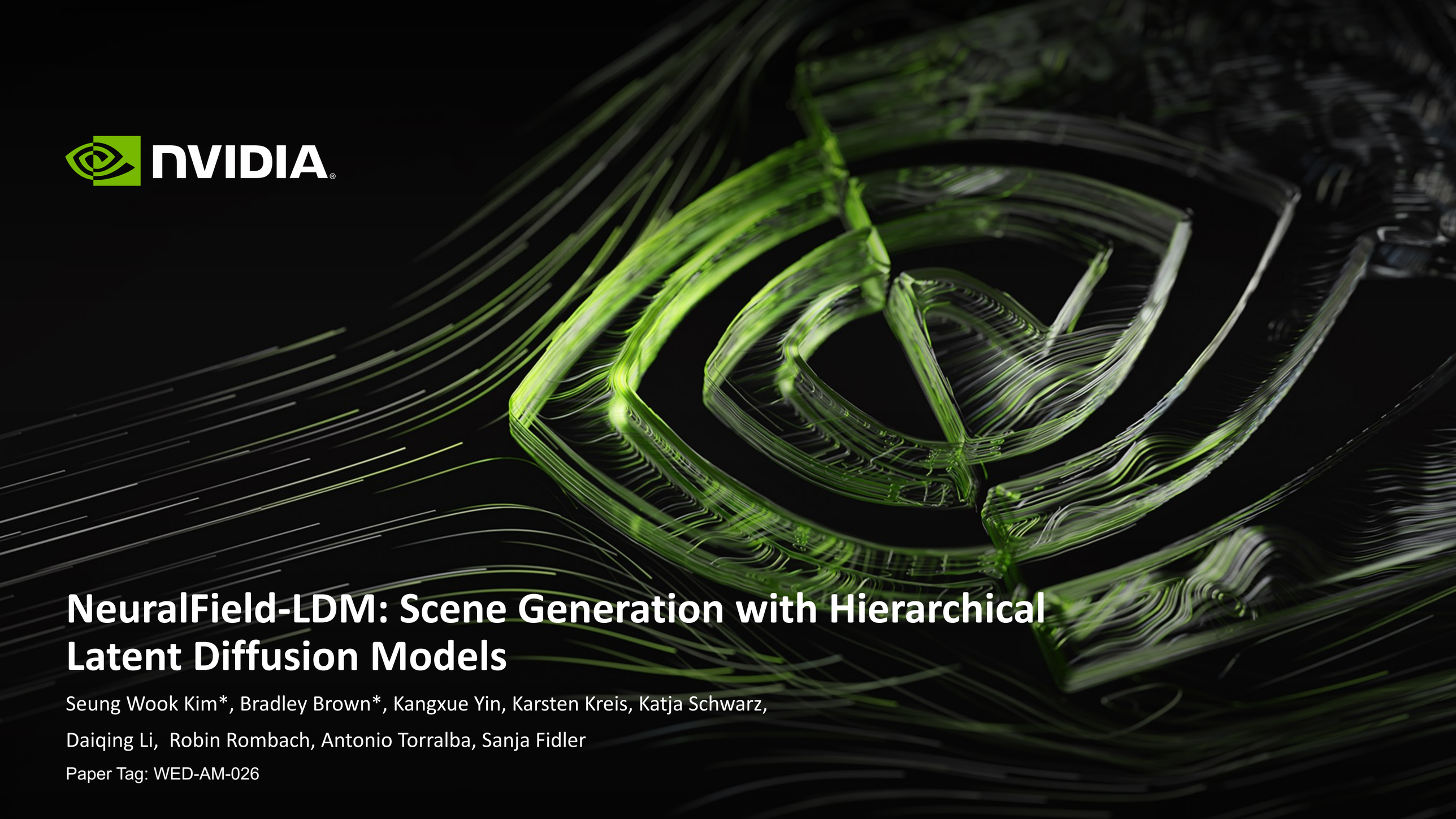


# NeuralField-LDM: Scene Generation with Hierarchical Latent Diffusion Models

Seung Wook Kim*, Bradley Brown*, Kangxue Yin, Karsten Kreis, Katja Schwarz,

Daiqing Li, Robin Rombach, Antonio Torralba, Sanja Fidler

Paper Tag: WED-AM-026

# Quick Preview: NeuralField-LDM

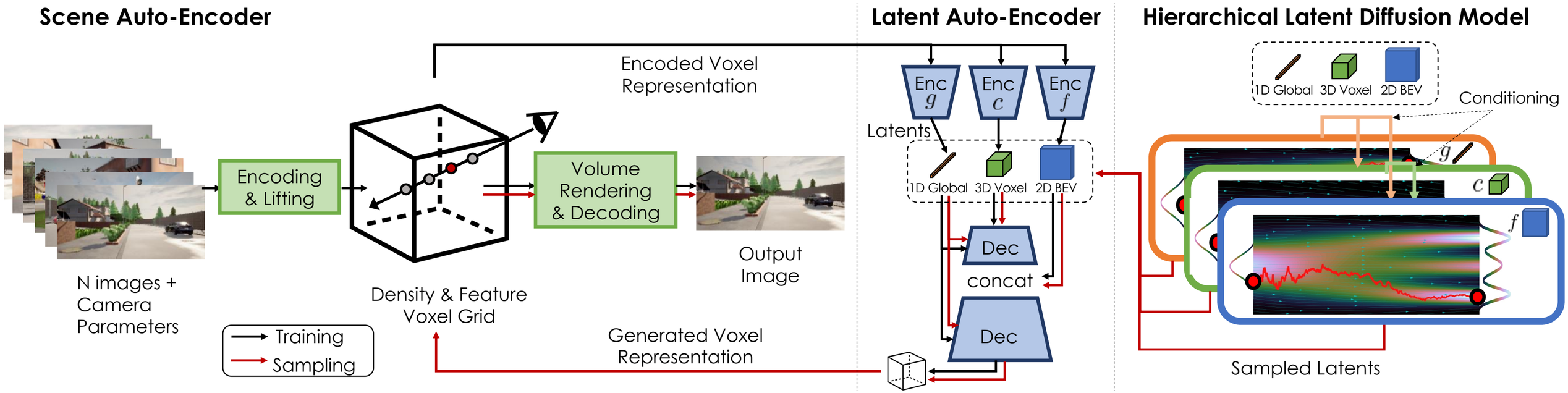
## Goal

- Given training data consisting of RGB images and their camera parameters, we aim to learn a ***3D scene generative model*** that can produce ***neural fields.***

# Quick Preview: NeuralField-LDM

## Overview

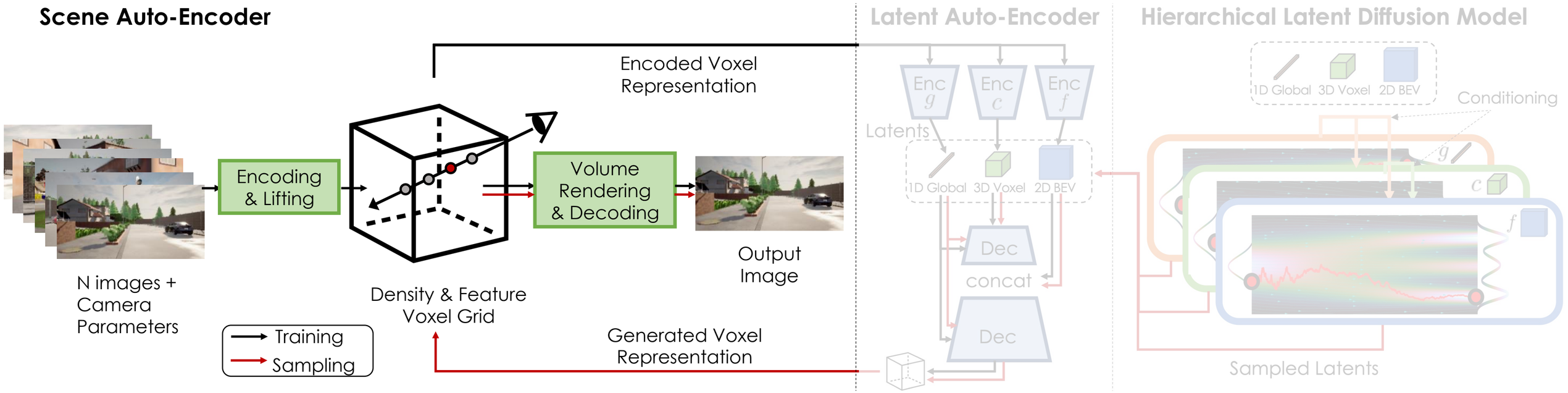
- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- Third Stage: Fit a hierarchical latent diffusion model on the latents.



# Quick Preview: NeuralField-LDM

## Overview

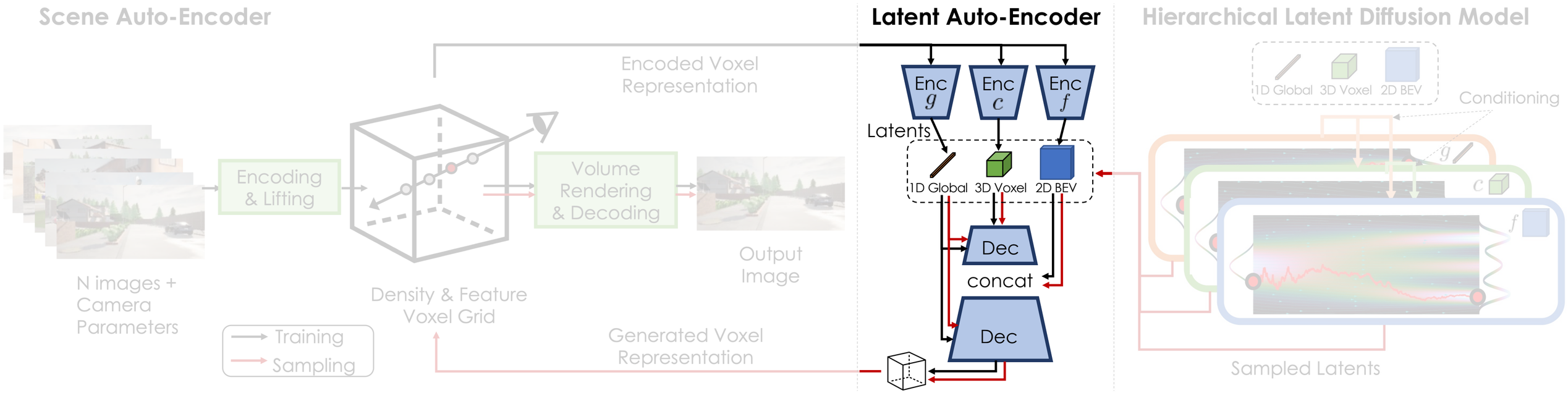
- **First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.**
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- Third Stage: Fit a hierarchical latent diffusion model on the latents.



# Quick Preview: NeuralField-LDM

## Overview

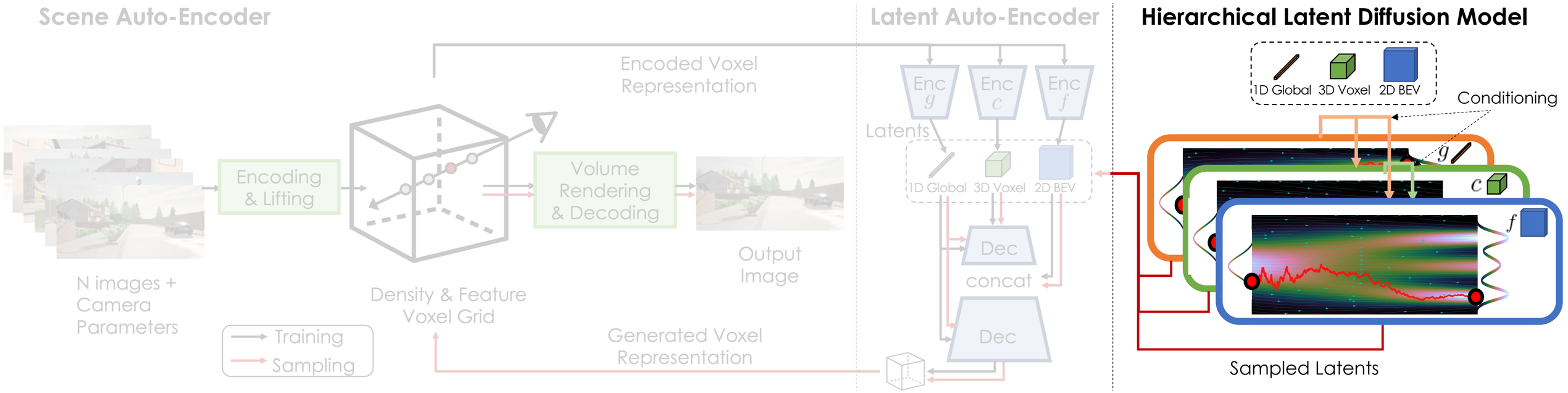
- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- **Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.**
- Third Stage: Fit a hierarchical latent diffusion model on the latents.



# Quick Preview: NeuralField-LDM

## Overview

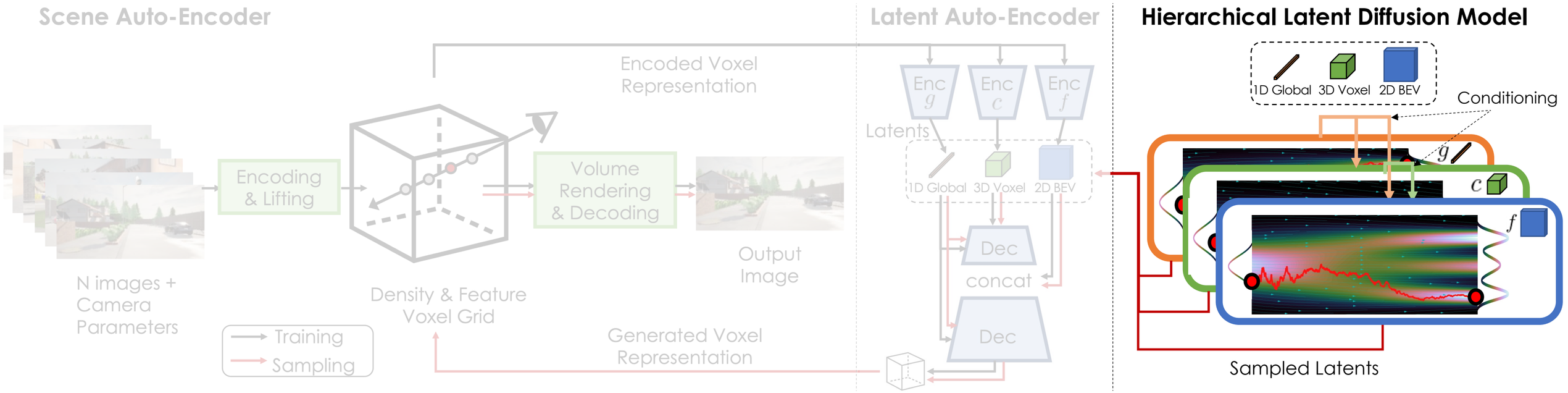
- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- **Third Stage: Fit a hierarchical latent diffusion model on the latents.**



# Quick Preview: NeuralField-LDM

## Overview

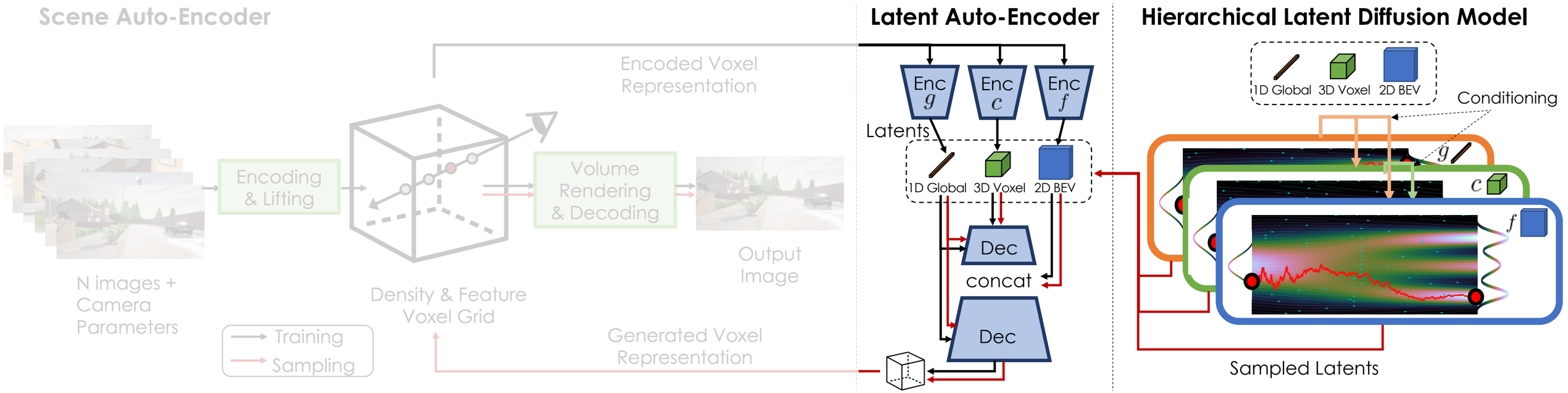
- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- Third Stage: Fit a hierarchical latent diffusion model on the latents.



# Quick Preview: NeuralField-LDM

## Overview

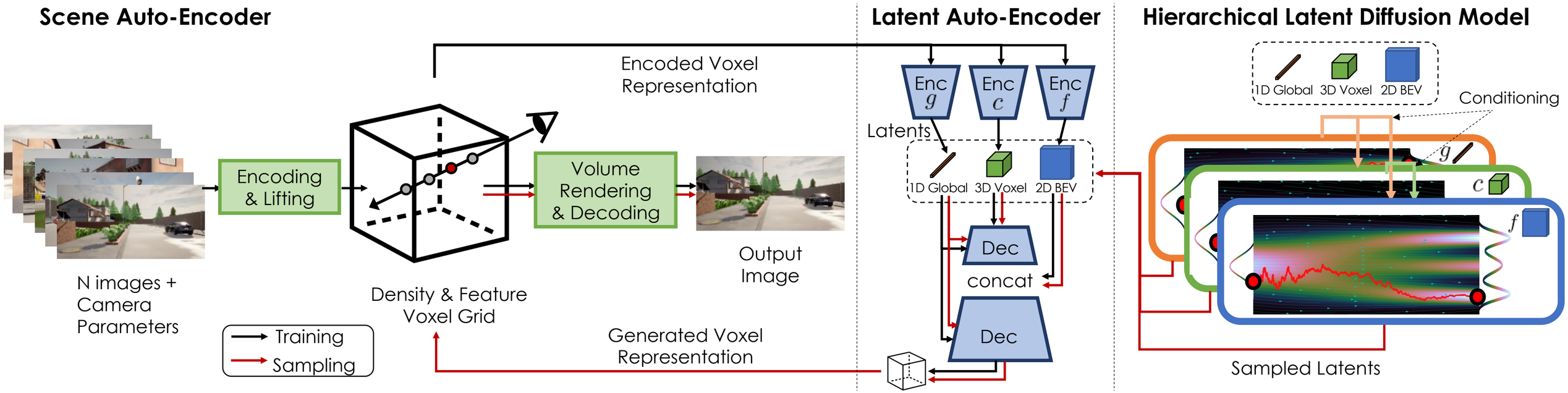
- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- Third Stage: Fit a hierarchical latent diffusion model on the latents.



# Quick Preview: NeuralField-LDM

## Overview

- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- Third Stage: Fit a hierarchical latent diffusion model on the latents.

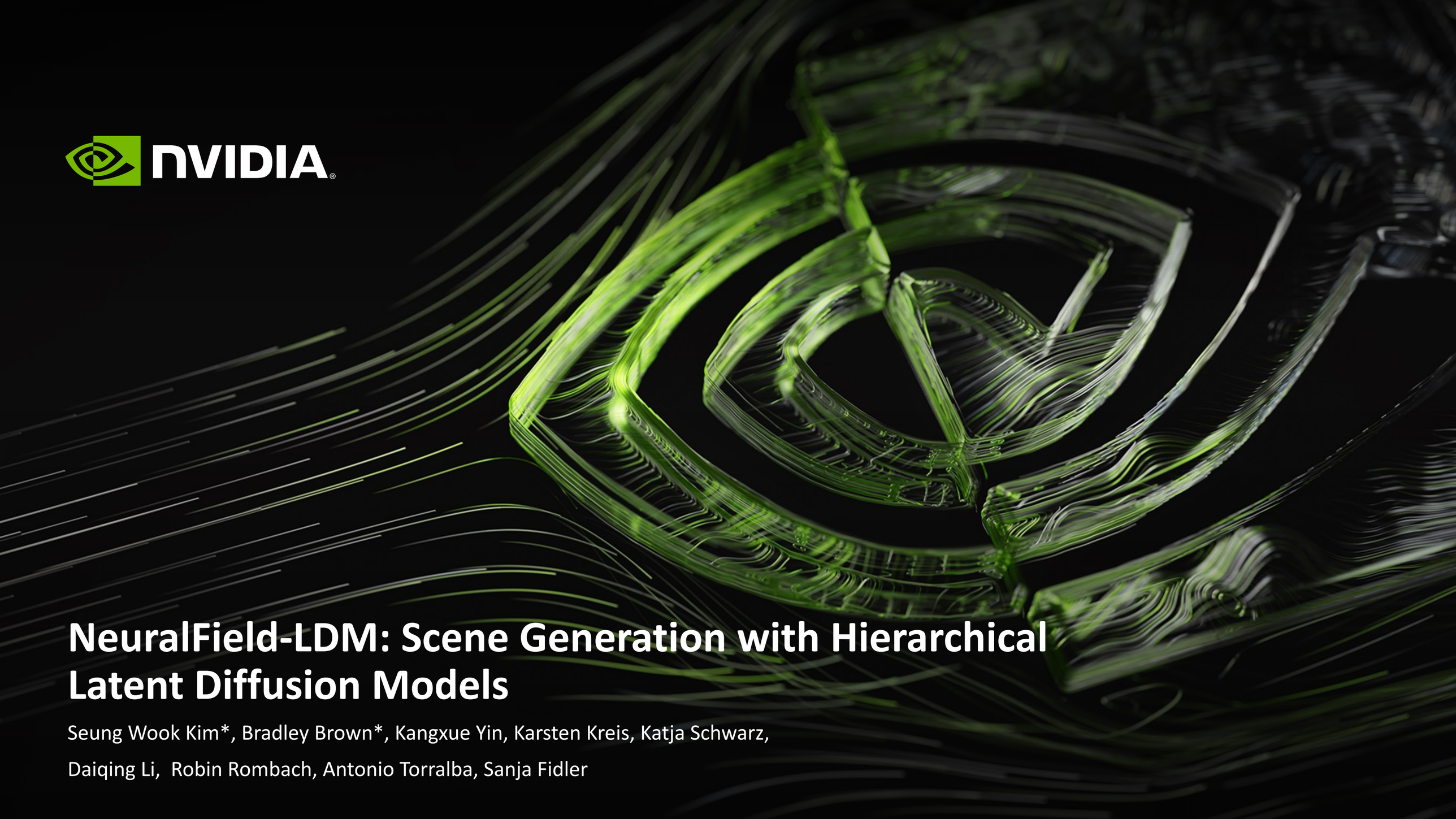


# Quick Preview: NeuralField-LDM

## Results – Generated Scenes

# Quick Preview: NeuralField-LDM

## Results – Stylized Scenes



# NeuralField-LDM: Scene Generation with Hierarchical Latent Diffusion Models

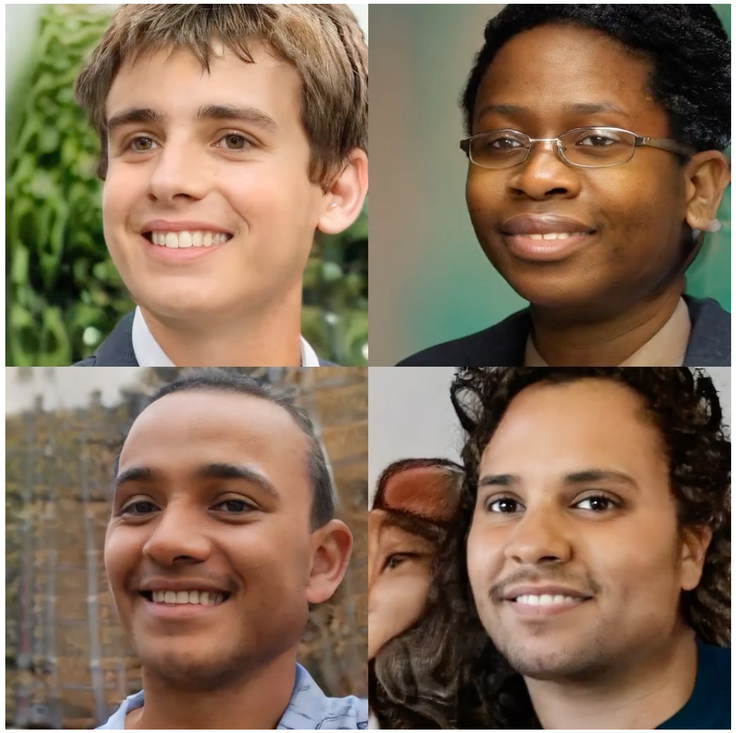
Seung Wook Kim*, Bradley Brown*, Kangxue Yin, Karsten Kreis, Katja Schwarz,

Daiqing Li, Robin Rombach, Antonio Torralba, Sanja Fidler

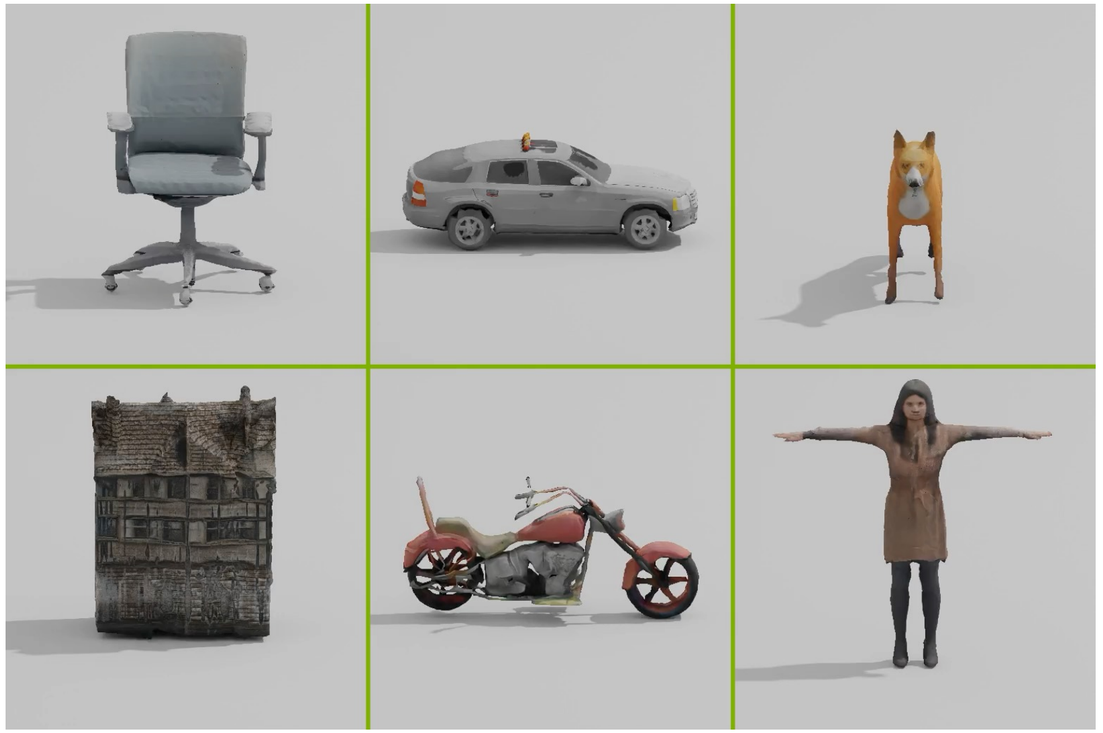
# 3D Generative Models

## Related Work

- Previous works mostly focus on object-level generation.



Chan et al., EG3D: Efficient Geometry-aware 3D Generative Adversarial Networks



Gao et al., GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images

# NeuralField-LDM

## Motivation

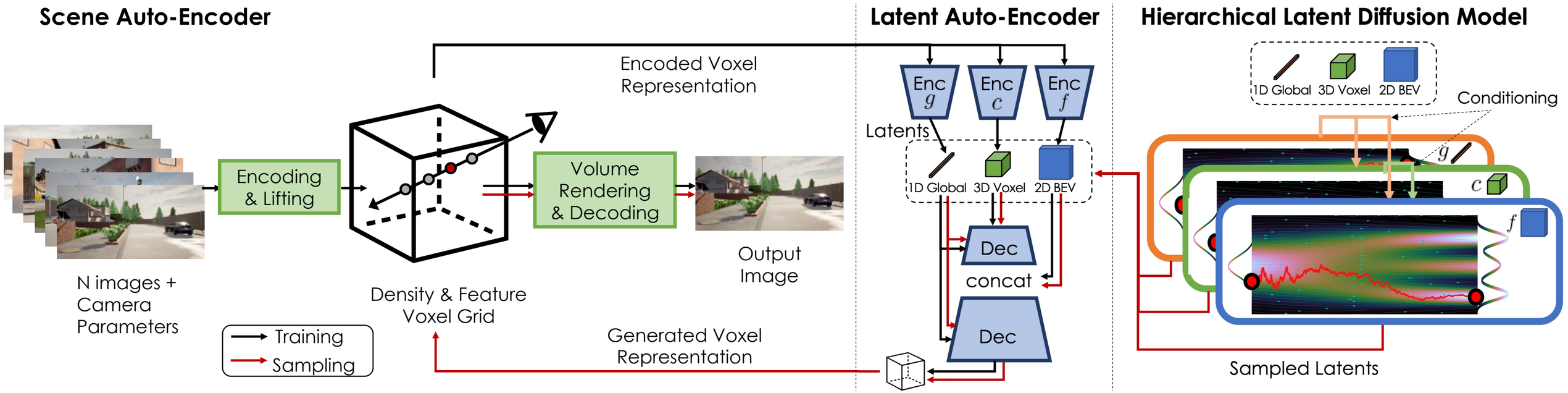
- How can we scale up to open-world scenes?
- Diffusion models are state-of-the-art generative models and admit powerful editing techniques.
- Acquiring large-scale ground-truth 3D texture and geometry data of the world is expensive.
- How can we train a diffusion model given training data in 2D video recordings, which are cheaper to obtain?


Source: https://twitter.com/zabsik_ua/status/1612823706862063616/photo/1

# NeuralField-LDM

## Overview

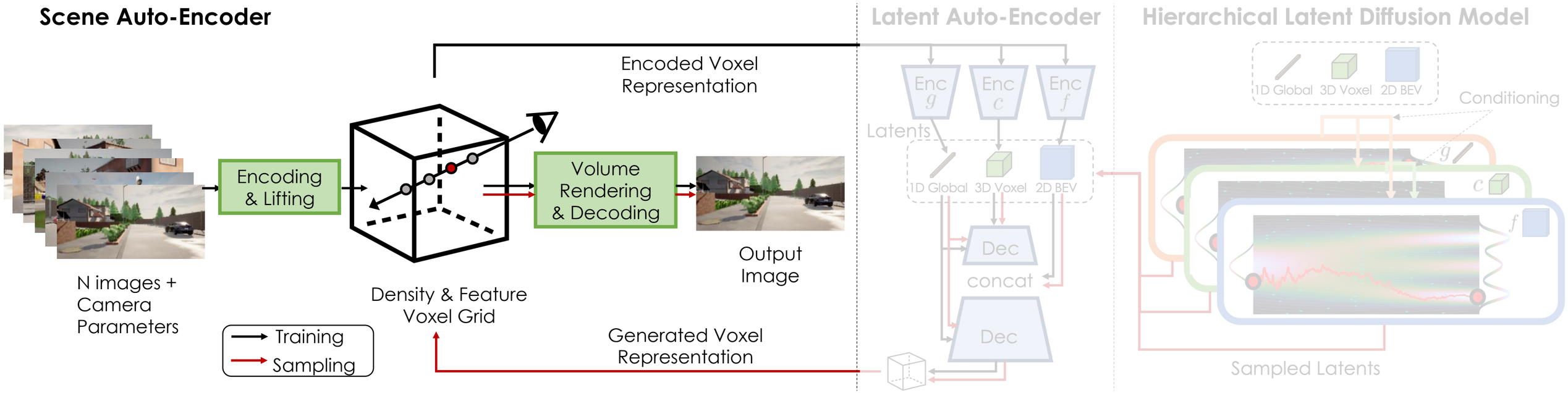
- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- Third Stage: Fit a hierarchical latent diffusion model on the latents.



# NeuralField-LDM

## Overview

- **First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.**
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- Third Stage: Fit a hierarchical latent diffusion model on the latents.



# First Stage: Scene Auto-Encoder

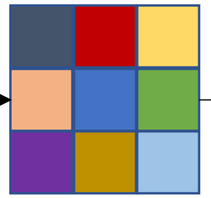
## Overview

Independently For Each Image

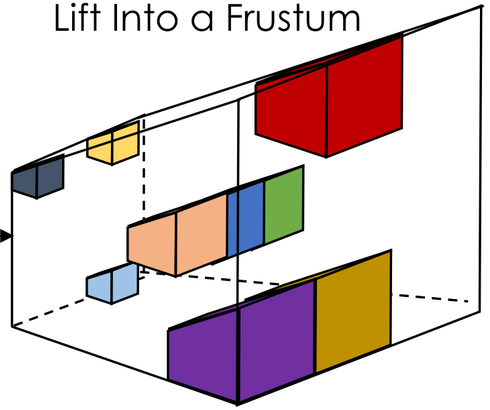
$i$

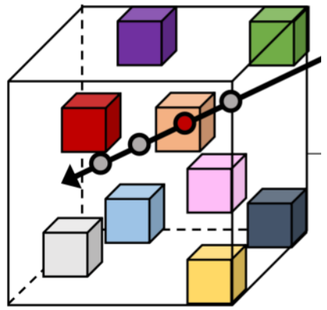
Encode & Get Features

2D CNN Encoder



Lift Into a Frustum





$V$

=

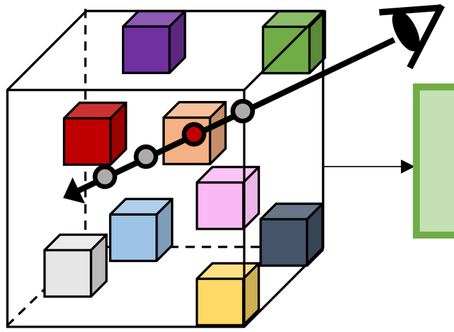
ZxYxXx1

$V_{\texttt{Density}}$

ZxYxXxC

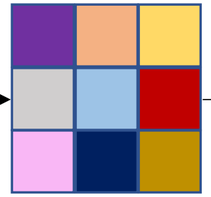
$V_{\texttt{Feat}}$

Merge The Frustums Across Different Views



Get 2D Features with Volume Rendering

Volume Rendering



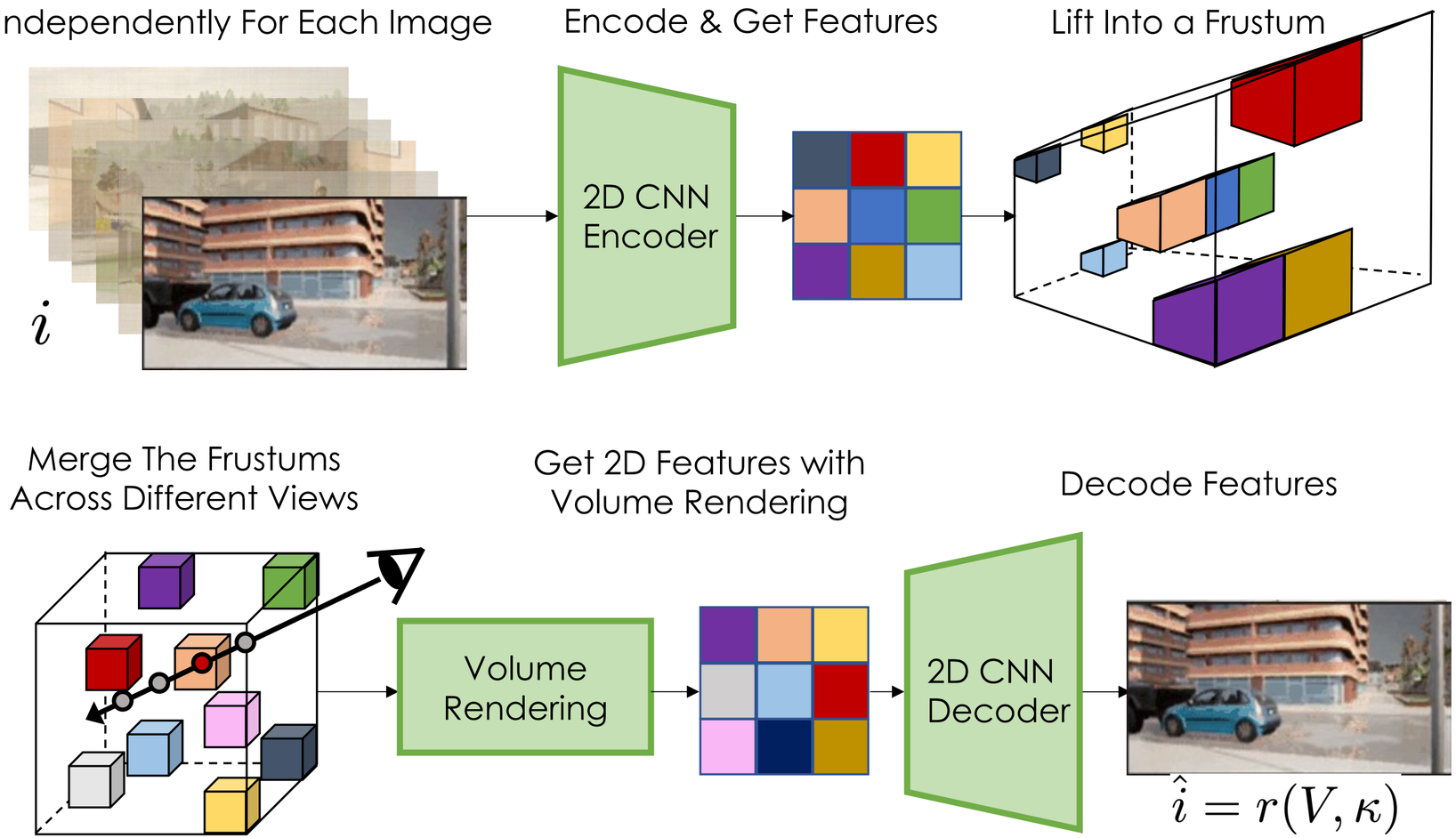
Decode Features

2D CNN Decoder

$\hat{i} = r(V, \kappa)$

# First Stage: Scene Auto-Encoder

## Overview




Losses:

- Reconstruction loss $||i - \hat{i}||$ on rendered images
- Depth loss $||\rho - \hat{\rho}||$ on volume-rendered depth
- Adversarial loss on the rendered images

# First Stage: Scene Auto-Encoder

## Novel View Synthesis

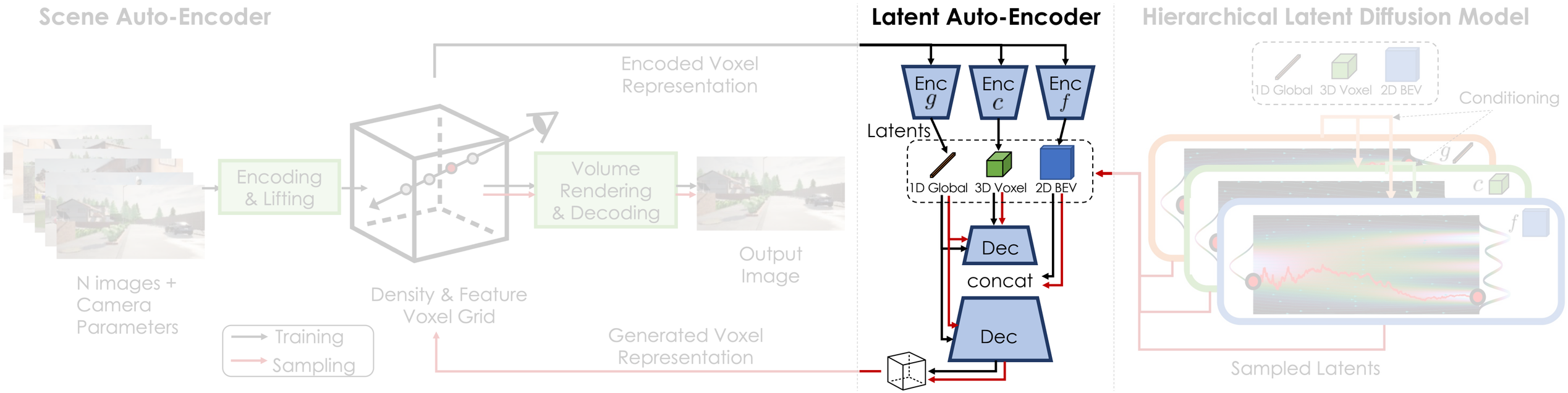
Input
Sequence

Novel view synthesis
from the encoded voxel

# NeuralField-LDM

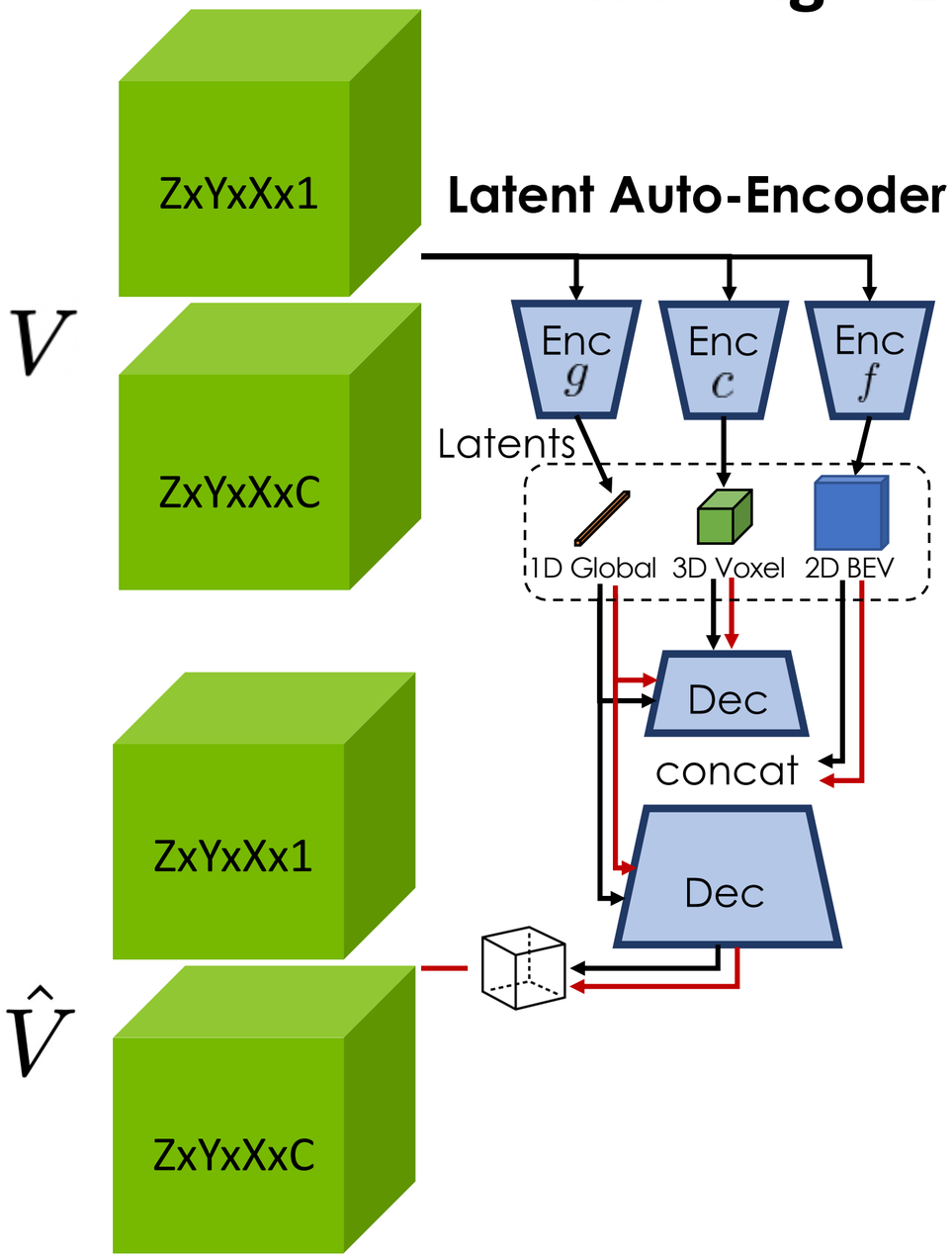
## Overview

- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- **Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.**
- Third Stage: Fit a hierarchical latent diffusion model on the latents.



# Second Stage: Latent Voxel Auto-Encoder

## Overview




- Difficult to fit a diffusion model on very high-dimensional data
- Latent Voxel Auto-Encoder compresses voxel grids to smaller latents
- Trained with voxel reconstruction loss

$$||V - \hat{V}||$$

and image reconstruction loss

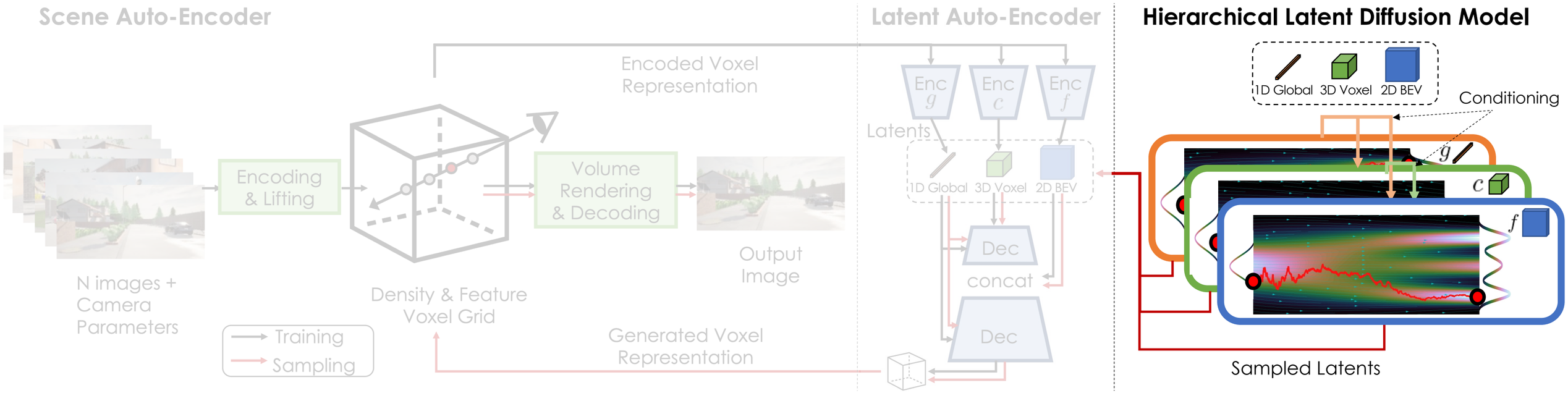
$$||i - \hat{i}||$$

where $\hat{i} = r(\hat{V}, \kappa)$

# NeuralField-LDM

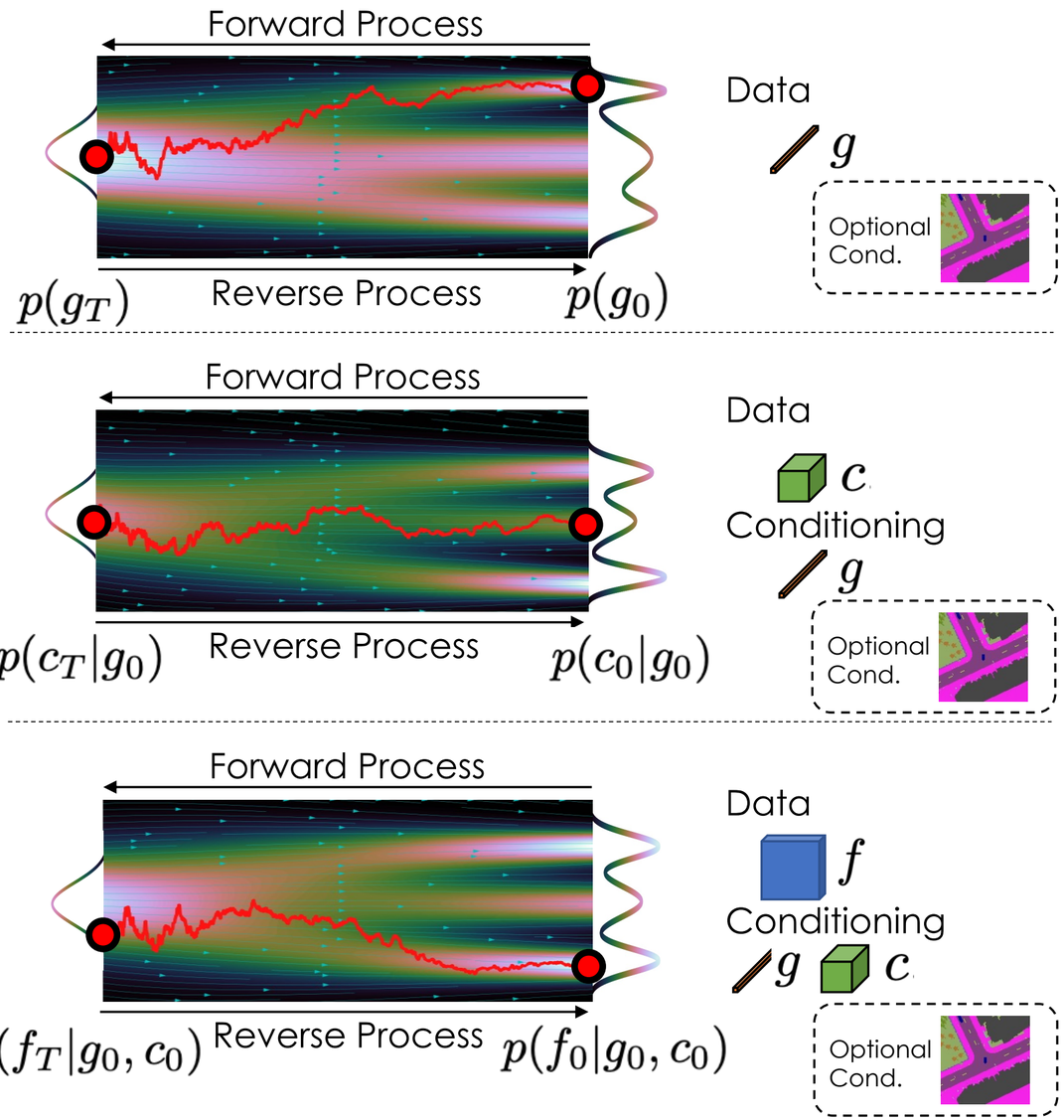
## Overview

- First Stage: Auto-encode multi-view input images into 3D density and feature voxel grids.
- Second Stage: Compress the feature volumes into smaller-dimensional latent spaces.
- **Third Stage: Fit a hierarchical latent diffusion model on the latents.**



# Third Stage: Hierarchical Latent Diffusion Models

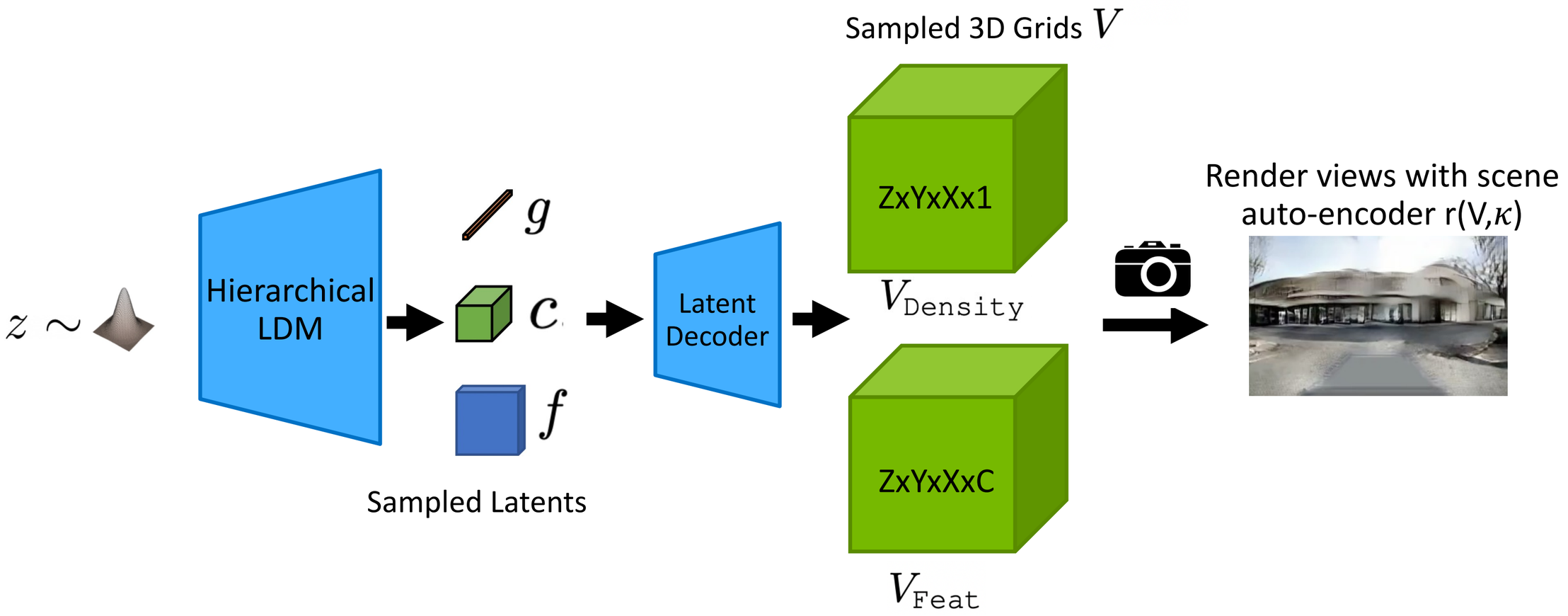
## Overview




- 1D Global latent $g$
  - Global properties of the scene, e.g. time of the day
- 3D Coarse latent $c$
  - Coarse 3D scene structure
- 2D Fine latent $f$
  - Further details for each location in bird's eye view perspective

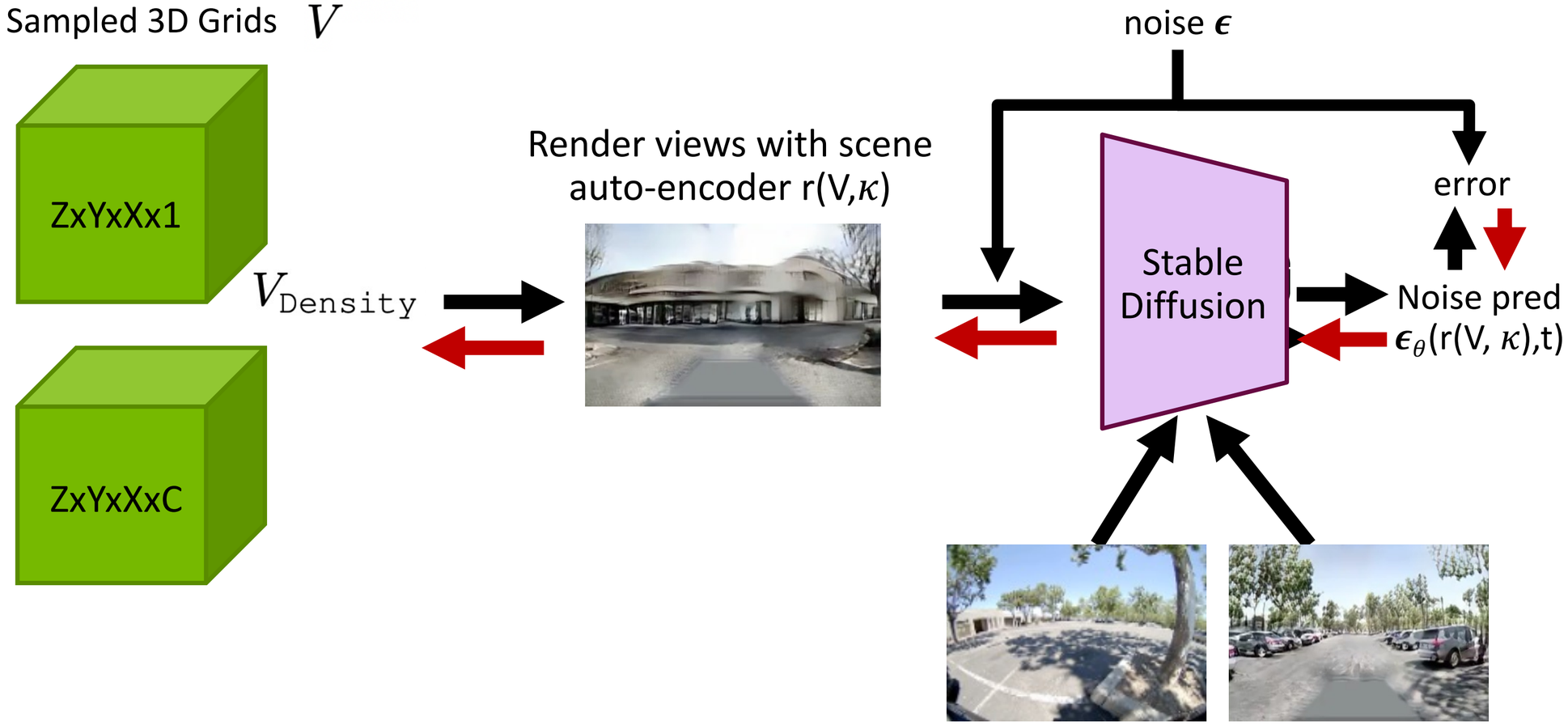
# NeuralField-LDM

## Sampling



# NeuralField-LDM

## Post-Optimization




$$\nabla_V L_{SDS} = \mathop{\mathbb{E}}_{\epsilon,t,\kappa}\left[w(\lambda_t)(\epsilon - \hat{\epsilon}_\theta(r(V,\kappa),t))\frac{\partial r(V,\kappa)}{\partial V}\right]$$



Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. arXiv, 2022.

# NeuralField-LDM

## Post-Optimization Conditioning

Render Conditioning
y’

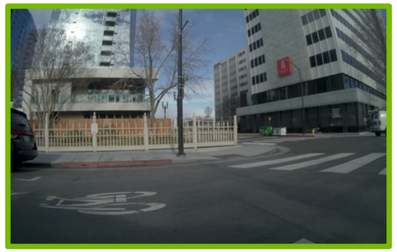
Stable Diffusion

Images with artifacts

Dataset Conditioning
y

Stable Diffusion

Clean images



Negative Guidance:

$$\gamma \nabla_x \log p(x|y) + (1-\gamma)\nabla_x \log p(x|y')$$

$$\frac{p(x|y)^\alpha}{p(x|y')}$$

# Scene Generation

## Results

Carla (Synthetic Simulator)

AVD (Human Driving Recordings)

# Scene Generation

## Results

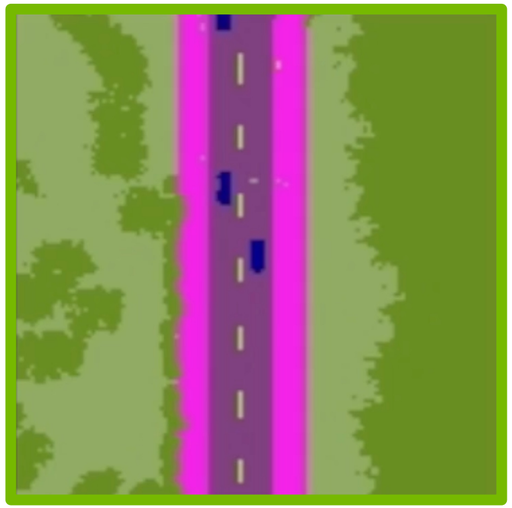
Generated Scenes

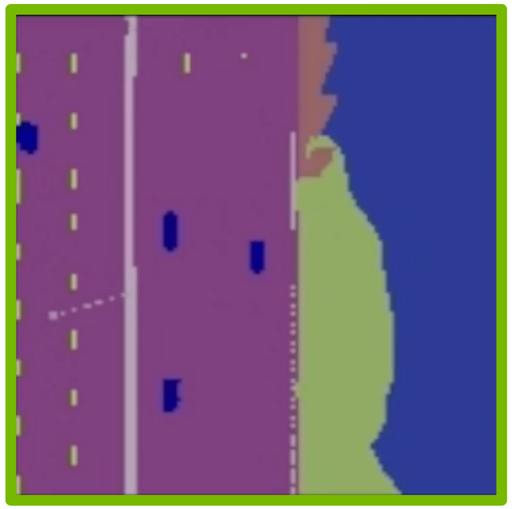
Finetuned with SDS Post-optimization

# Conditional Generation

## Results

Bird’s Eye View Semantic Segmentation Map





# Stylizing Generated 3D Scenes

## Results

Generated 3D Scene

Winter Wonderland

Medieval Castle

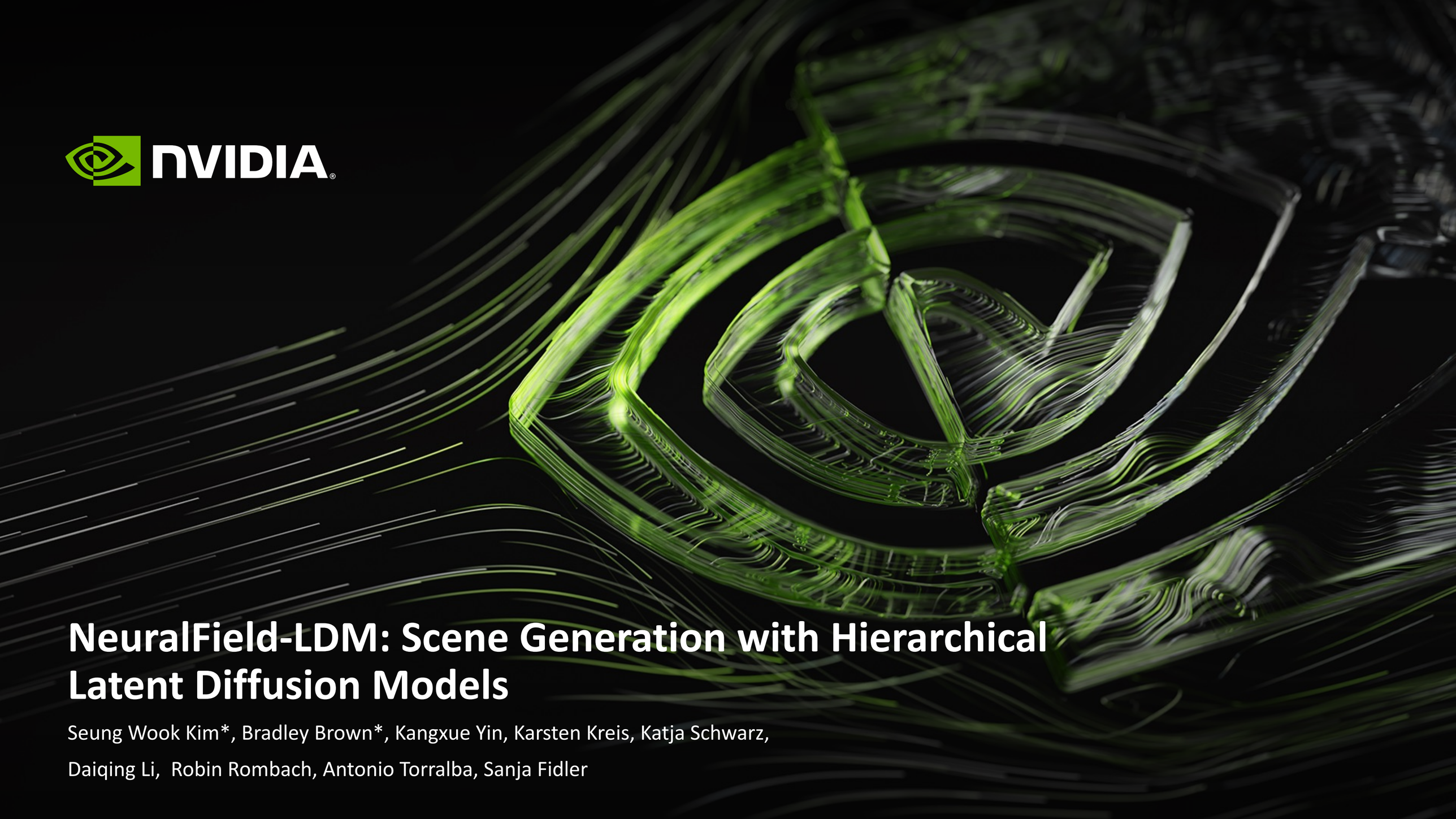
# Scene Editing

## Results

Generated 3D Scene

Stylized Scene

Combine voxels



# NeuralField-LDM: Scene Generation with Hierarchical Latent Diffusion Models

Seung Wook Kim*, Bradley Brown*, Kangxue Yin, Karsten Kreis, Katja Schwarz,

Daiqing Li, Robin Rombach, Antonio Torralba, Sanja Fidler