

# LSRM: High-Fidelity Object-Centric Reconstruction via Scaled Context Windows

Zhengqin Li<sup>1,2</sup>, Cheng Zhang<sup>1</sup>, Jakob Engel<sup>1</sup>, and Zhao Dong<sup>1</sup>

<sup>1</sup>Meta Reality Labs Research, USA    <sup>2</sup>NVIDIA, USA

**Abstract.** We introduce the **L**arge **S**parsity **R**econstruction **M**odel to study how scaling transformer context windows affects feed-forward 3D reconstruction. Although recent object-centric feed-forward methods produce robust, high-quality reconstructions, they still lag behind dense-view optimization in recovering fine-grained texture and appearance. We show that expanding the context window—by substantially increasing the number of active object and image tokens—narrows this gap and enables high-fidelity 3D object reconstruction and inverse rendering. To scale effectively, we adapt native sparse attention [68] for 3D reconstruction with three key contributions: (1) an efficient coarse-to-fine pipeline that focuses computation on informative regions by predicting sparse high-resolution residuals; (2) a 3D-aware spatial routing mechanism that establishes accurate 2D-3D correspondences using explicit geometric distances rather than standard attention scores; and (3) a custom block-aware sequence-parallel strategy with an All-gather-KV protocol to balance dynamic, sparse workloads across GPUs. As a result, LSRM handles **20×** more object tokens and **>2×** more image tokens than prior state-of-the-art (SOTA) methods. Extensive evaluations on standard novel-view synthesis benchmarks show substantial gains over the current SOTA, yielding **>2.4** dB higher PSNR and **>40%** lower LPIPS. Furthermore, when extending LSRM to inverse rendering, qualitative and quantitative evaluations on widely used benchmarks demonstrate consistent improvements in texture and geometry details, achieving an LPIPS that matches or exceeds that of SOTA dense-view optimization methods. Code and model weights are available on our project page.

**Keywords:** Object-centric feed-forward reconstruction · Sparse attention · 3D foundation model

## 1 Introduction

Recent years have witnessed rapid progress in the application of 3D foundation models—typically built on large-scale transformer architectures [55]—to tackle 3D tasks previously considered intractable. These tasks include joint estimation of geometry and camera parameters [30, 44, 56, 57], dynamic scene reconstruction [41, 43, 46, 64, 72], and sparse-view reconstruction [20, 24, 58, 73, 78] and inverse rendering [38, 69]. In object-centric reconstruction and inverse rendering



**Fig. 1: High-fidelity 3D reconstruction.** Given 6–18 images (**left**), LSRM adapts Native Sparse Attention (NSA) to generate explicit meshes and textures in a single feed-forward pass with minimal compute overhead. As an extension, LSRM can also predict BRDF maps for inverse rendering (**right**). Zoom in for details.

in particular, feed-forward models cut reconstruction time by orders of magnitude compared with optimization-based pipelines [48, 52, 74, 75], while achieving competitive view synthesis and relighting quality, and often exhibiting stronger geometric robustness on challenging specular objects [38]. Yet a key limitation persists: reconstructed textures frequently miss fine-grained detail, leading to blurry text, smeared logos, and distorted facial features. Recent attempts to address this—by swapping the underlying 3D representation [24, 73] or adding post-hoc texture refinement [50]—improve results but still do not match the fidelity of dense-view optimization. Narrowing this gap is essential for deploying feed-forward 3D reconstruction at the quality and reliability required for practical, industrial-scale 3D digital twin creation.

We hypothesize that this gap can be narrowed by expanding the context window—specifically, by increasing the number of active object and image tokens. However, naïvely upscaling the resolution of 3D object or 2D image representations rapidly becomes impractical, incurring quadratic (and in volumetric settings, cubic) growth in GPU memory and compute. To overcome this bottleneck, we adapt native sparse attention (NSA) [68], a hardware-friendly algorithm that restricts each token’s attention to a sparse set of token blocks, enabling efficient training and inference at substantially larger context sizes. Building on NSA, we develop an efficient coarse-to-fine training pipeline. In the first stage, a dense reconstruction transformer identifies the most informative volume tokens; in parallel, we prune 2D image tokens by retaining only those covering foreground pixels. This coarse model provides a strong initialization for both the network weights and the geometric structure. In the second stage, the sparse re-

construction transformer refines the reconstruction by predicting high-resolution sparse volume residuals, an approach we find considerably more effective than directly regressing a high-resolution volume from scratch.

While enlarging the context window with NSA already improves baseline quality, we introduce two critical architectural adaptations to further maximize reconstruction fidelity and hardware efficiency. First, we observe that the default NSA block selection [68]—which relies purely on attention scores from compressed tokens—often fails to capture reliable correspondences between voxel and image tokens, especially in early transformer layers. To remedy this, we propose a 3D-aware spatial routing strategy that explicitly uses the geometric distances established in Stage 1, alongside known camera parameters, to retrieve relevant blocks. This 3D-aware routing markedly enhances texture sharpness, rendering previously blurred fine text legible. Second, the dynamic sparse layout naturally induces severe token imbalance across GPUs. To address this, we design a custom block-aware sequence parallelism scheme. We shard tokens by their spatial 2D and 3D blocks, ensuring tokens within the same block remain localized on the same GPU to eliminate cross-device communication during local attention. Furthermore, we leverage the high KV compression of Grouped-Query Attention (GQA) [1] to implement an All-gather-KV protocol. This efficiently broadcasts a lightweight KV cache across the node, providing full global context with minimal communication overhead. Together, these designs enable LSRM to process  $20\times$  more object tokens and  $>2\times$  more image tokens than the prior SOTA [38], without a proportional increase in computational resources. Comprehensive experiments on standard object-centric 3D reconstruction benchmarks demonstrate substantial gains, yielding  $>2.4$  dB higher PSNR and a  $>40\%$  reduction in LPIPS on the GSO dataset [14]. We further extend LSRM to the task of inverse rendering by fine-tuning the model to predict material reflectance from images captured under natural illumination. Both qualitative and quantitative experiments show significant improvements in textural and geometric detail. Specifically on the widely used StanfordORB dataset [27], LSRM sets a new standard for feed-forward inverse rendering methods, achieving an LPIPS that matches the SOTA dense-view optimization method [52].

## 2 Related Work

*Object-Centric Feed-Forward Reconstruction* 3D foundation models trained on large-scale datasets [11] have driven major progress in object-centric reconstruction. Pioneering models like LRM [20] achieved impressive single-image geometry recovery, but they often struggle to reproduce high-fidelity textures. Subsequent works have sought to improve appearance quality through architectural refinements [58], alternative 3D representations [7, 24, 63, 73, 83], progressive updates [38], and post hoc texture refinement [50]. Despite these advances, a sizable performance gap compared to dense-view, optimization-based pipelines persists. A recent approach [77] narrows this gap via test-time training to effectively enlarge the context window; however, it lacks an explicit 3D representation and may

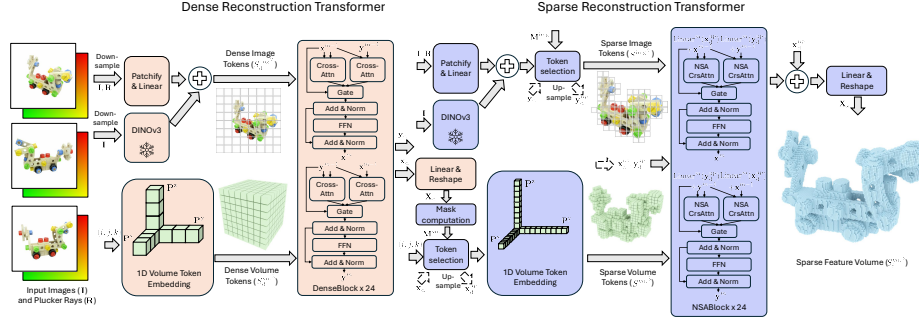
suffer from view synthesis degradation at unfamiliar camera angles. In contrast, LSRM leverages scaled context windows to explicitly reconstruct high-fidelity 3D geometry and textures in a single forward pass. Furthermore, by extending LSRM to predict BRDF material maps, we enable seamless downstream editing and realistic relighting within standard graphics pipelines.

*Inverse Rendering* Inverse rendering aims to decompose images into their intrinsic factors—geometry, materials, and illumination—so that scenes or objects can be edited and re-rendered realistically. Classical measurement-based approaches [2, 10, 17] rely on specialized capture rigs and controlled environments to densely sample view and lighting directions. Recent optimization-based methods [3–5, 15, 19, 23, 25, 42, 48, 49, 52, 66, 74, 75, 79–81] can often work under natural illumination, but they usually require dense multi-view inputs and remain fragile in ill-posed conditions such as saturated specularities, cast shadows, and complex interreflections. Many learning-based techniques focus on predicting per-pixel BRDF maps [6, 12, 33, 35–37, 39, 47, 70] without producing a complete 3D representation. Recently, propelled by advances in object-centric 3D reconstruction models, pioneering works [38, 69] have shown that feed-forward models can output fully relightable 3D objects with explicit geometry and materials. LSRM builds on this direction and further improves reconstruction and appearance fidelity by enabling substantially longer context windows.

*Sparsity in 3D Generation and Reconstruction* Many prior approaches handle the inherent sparsity of 3D surfaces indirectly, e.g., by compressing geometry into compact representations [18, 21, 28, 31, 32, 34, 59, 71, 76, 82] or adopting localized point-based formats such as 3D Gaussians [9, 29, 54, 65]. More recently, a line of 3D generative work has modeled sparsity explicitly through structured latent representations—such as sparse voxel grids—to scale volumetric resolution substantially [40, 60–62]. In these frameworks, sparse volumes are either encoded by sparse 3D VAEs into dense latent codes [8, 40, 61, 62] or, as in Direct3D-S2 [60], downsampled to a lower-resolution sparse volume that can be generated with an NSA-style transformer operating over 3D token blocks. While these methods demonstrate efficiency in processing high-resolution volumes, their focus remains restricted to synthesizing intricate geometric details, often neglecting texture. Even when texture generation is supported [8, 61], the resulting texture fidelity remains limited even under clean synthetic inputs. Such single-view generative models target plausible 3D appearance synthesis rather than high-fidelity deterministic reconstruction from posed multi-view inputs, making them complementary to our setting. This highlights that high-fidelity textured object reconstruction is a distinct and challenging problem, motivating the specialized architecture design by LSRM.

### 3 Method

This section details the architecture of LSRM. We first formulate the Stage 1 dense reconstruction transformer, specifying the model’s inputs, outputs, and



**Fig. 2: Network architecture of LSRM.** Our method employs a two-stage coarse-to-fine pipeline. In Stage 1, a Dense Reconstruction Transformer generates a coarse, low-resolution volume. Stage 2 utilizes this coarse volume to initialize the active sparse volume tokens and predicts high-resolution sparse residuals, constructing the final 3D representation.

notation. We then describe our Stage 2 sparse reconstruction transformer, which is initialized from the dense model and adapts NSA to significantly expand the context window. Finally, we detail two key improvements that maximize reconstruction fidelity and hardware efficiency. Additional background on NSA is provided in the supplementary material.

### 3.1 Stage 1: Coarse Dense Reconstruction for Initialization

We first train a dense reconstruction transformer for lower resolution reconstruction; its weights and intermediate predictions are then used to initialize the Stage 2 sparse reconstruction transformer.

*Image and Volume Tokenization* The inputs consist of a sparse set of  $M$  posed images ( $M \in [12, 18]$  similar to [38]), with camera parameters explicitly encoded as Plücker rays. Let  $\{\mathbf{I}_m\}_{m=1}^M$  denote the non-overlapping  $8 \times 8$  patches extracted from the  $256 \times 256$  input views, yielding a spatial token map of resolution  $S_d^{\text{img}} = 32$  per view. Let  $\{\mathbf{R}_m\}_{m=1}^M$  denote their corresponding Plücker rays. We compute the combined multi-view image tokens,  $\mathbf{y}$ , by fusing a linear projection of the patches and rays with features extracted from a frozen DINOv3 [51] encoder:

$$\mathbf{y} = \{\text{Linear}(\mathbf{I}_m, \mathbf{R}_m) + \text{Linear}(\text{DINOv3}(\text{Upsample}(\mathbf{I}_m)))\}_{m=1}^M \quad (1)$$

Because DINOv3 natively operates on  $16 \times 16$  patches, the input images are upsampled by 2 prior to encoding. Incorporating DINOv3 features significantly accelerates training convergence and improves the model’s generalizability.

The volume tokens  $\mathbf{x}$  are initialized from learned 3D positional embeddings. To avoid wasting GPU memory when later scaling to high-resolution grids in the Stage 2 sparse transformer, we factorize the embeddings along the three spatial

axes. Instead of learning a dense parameter volume, we learn three independent 1D embeddings:  $\mathbf{P}^x \in \mathbb{R}^{S_d^{\text{vol}} \times d}$ ,  $\mathbf{P}^y \in \mathbb{R}^{S_d^{\text{vol}} \times d}$ , and  $\mathbf{P}^z \in \mathbb{R}^{S_d^{\text{vol}} \times d}$ , where  $d = 1024$ ,  $S_d^{\text{vol}} = 16$  are the embedding dimension and the dense token volume’s resolution. For a token at spatial coordinate  $(i, j, k)$ , its embedding  $\mathbf{p}_{i,j,k}$  is simply the element-wise sum:

$$\mathbf{p}_{i,j,k} = \mathbf{P}_i^x + \mathbf{P}_j^y + \mathbf{P}_k^z \quad (2)$$

The resulting initialized volume tokens,  $\mathbf{x} = \{\mathbf{p}_{i,j,k}\}$ , are then concatenated with image tokens  $\mathbf{y}$  and processed by the dense transformer blocks.

*Dense Reconstruction Transformer* Following prior works [38, 58], our dense transformer consists of 24 blocks with hidden dimension 1024. Unlike prior dense models, we use grouped-query attention with 32 query heads and only 2 key/value heads, providing the high KV-head compression required by the subsequent sparse transformer and exploited by our sequence-parallel training (Sec. 3.4). Additionally, rather than standard monolithic self-attention, we use a gated mixture of cross-attention updates to evolve token representations. This decoupled design provides flexible, fine-grained control over attention branches in Stage 2—for example, allowing independent tuning of the number of selected KV-blocks ( $L$ ) for image and volume tokens. Let  $\mathbf{x}$  denote volume tokens and  $\mathbf{y}$  denote image tokens. The attention module within one DenseBlock is:

$$\mathbf{o}^x = \mathbf{w}^{x\text{-self}} \text{CrossAttn}(\mathbf{x}, \mathbf{x}) + \mathbf{w}^{x\text{-cross}} \text{CrossAttn}(\mathbf{x}, \mathbf{y}) \quad (3)$$

$$\mathbf{o}^y = \mathbf{w}^{y\text{-self}} \text{CrossAttn}(\mathbf{y}, \mathbf{y}) + \mathbf{w}^{y\text{-cross}} \text{CrossAttn}(\mathbf{y}, \mathbf{x}) \quad (4)$$

$$\mathbf{w}^{x\text{-self}}, \mathbf{w}^{x\text{-cross}} = \text{Sigmoid}(\text{Linear}_{\text{gate}}^x(\mathbf{x})) \quad (5)$$

$$\mathbf{w}^{y\text{-self}}, \mathbf{w}^{y\text{-cross}} = \text{Sigmoid}(\text{Linear}_{\text{gate}}^y(\mathbf{y})) \quad (6)$$

We denote by  $\mathbf{x}_d$  and  $\mathbf{y}_d$  the volume and image token outputs from the final DenseBlock. We decode the volume tokens  $\mathbf{x}_d$  into a dense feature volume via a linear projection. This layer upsamples spatial resolution by  $4\times$ , producing a grid of size  $S_{\text{df}}^{\text{vol}} = 64$  while reducing feature dimension to  $d_f = 32$ :

$$\mathbf{X}_d = \text{Linear}(\mathbf{x}_d), \quad \mathbf{X}_d \in \mathbb{R}^{S_{\text{df}}^{\text{vol}^3} \times d_f} \quad (7)$$

This dense feature volume,  $\mathbf{X}_d$ , can be utilized for either novel view synthesis or explicit textured mesh extraction via lightweight MLP decoders. With a slight abuse of notation, let  $\mathbf{p} \in \mathbb{R}^3$  be a continuous 3D point sampled within the volume. We compute its corresponding properties as follows:

$$\mathbf{f} = \text{Trilinear}(\mathbf{X}_d; \mathbf{p}) \quad (8)$$

$$\mathbf{z} = \text{Sigmoid}(\text{MLP}_{\mathbf{z}}(\mathbf{f})), \quad \mathbf{z} \in \{\mathbf{a}, \mathbf{r}, \mathbf{m}\} \text{ or } \mathbf{z} = \mathbf{c} \quad (9)$$

$$s = \text{MLP}_s(\mathbf{f}) + s_{\text{bias}}(\mathbf{p}) \quad (10)$$

where  $\mathbf{f}$  is the trilinearly interpolated feature vector,  $s$  is the predicted SDF, and  $s_{\text{bias}}$  is an effective offset adapted from [38].  $\mathbf{c}$  denotes the emitted color for novel view synthesis. Alternatively  $\mathbf{a}$ ,  $\mathbf{r}$ , and  $\mathbf{m}$  represent the albedo, roughness, and metallic material properties for inverse rendering. For rendering, we adopt the VolSDF [67] formulation due to its simplicity and effectiveness.

### 3.2 Stage 2: Sparse Reconstruction with Long Context Windows

For high-fidelity reconstruction in Stage 2, we substantially increase the resolutions of both the input images (e.g.,  $S^{\text{img}} = 3S_d^{\text{img}} = 96$ ) and the volume grid (e.g.,  $S^{\text{vol}} = 6S_d^{\text{vol}} = 96$ ). To efficiently handle the resulting expanded token sequence, we first describe how we extract active, spatially-sparse token subsets, and then present the spatial block partitioning required by our NSA mechanism, followed by the architecture of the sparse reconstruction transformer.

*Informative Token Selection* We build a spatially-sparse token representation by retaining only the most informative image and volume tokens. For image tokens, we discard the background, preserving only patches that contain foreground pixels based on the foreground mask  $\mathbf{M}^{\text{img}}$ . For the volume tokens, we leverage the geometric prior from the Stage 1 dense transformer to compute a binary mask  $\mathbf{M}^{\text{vol}}$ . This mask identifies and retains only the voxels near the object surface, as these primarily determine the final appearance. Concretely, within a voxel at spatial coordinate  $(i, j, k)$ , we uniformly sample a grid of  $T = 4^3 = 64$  points  $\mathcal{P}_{i,j,k} = \{\mathbf{p}_t\}_{t=1}^T$ . We evaluate the SDF value  $s_t$  at each point using the frozen Stage 1 dense feature volume  $\mathbf{X}_d$  and set the voxel mask  $\mathbf{M}_{i,j,k}^{\text{vol}}$  based on a distance threshold  $\tau$ :

$$s_t = \text{MLP}_s(\text{Trilinear}(\mathbf{X}_d; \mathbf{p}_t))$$

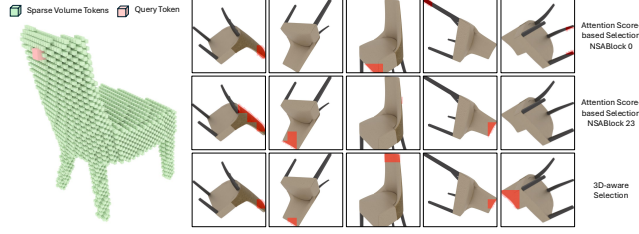
$$\mathbf{M}_{i,j,k}^{\text{vol}} = \mathbb{I}\left(\min_{\mathbf{p}_t \in \mathcal{P}_{i,j,k}} |s_t| \leq \tau \quad \text{or} \quad \left(\min_{\mathbf{p}_t \in \mathcal{P}_{i,j,k}} s_t\right) \cdot \left(\max_{\mathbf{p}_t \in \mathcal{P}_{i,j,k}} s_t\right) \leq 0\right) \quad (11)$$

where  $\mathbb{I}(\cdot)$  denotes the indicator function. This rule ensures that only voxels close to the underlying surface, or directly intersected by it, are instantiated as tokens for high-resolution processing.

*Block Partitioning and Compression* A naïve 1D sequence partitioning of tokens destroys spatial locality, which is detrimental to fine-grained 3D reconstruction. Therefore, we adopt a spatial block partitioning strategy similar to Direct3D-S2 [60]. We group active image and volume tokens based on their original 2D and 3D coordinates. Specifically, space is divided into discrete blocks of size  $8 \times 8$  for images and  $8 \times 8 \times 8$  for the volume, reducing the block-level resolutions to  $S_b^{\text{img}} = S^{\text{img}}/8$  and  $S_b^{\text{vol}} = S^{\text{vol}}/8$ . Because uninformative tokens are removed, each spatial block contains a variable number of active tokens. To compute the block-level compressed keys ( $\mathbf{k}^{\text{cmp}}$ ) and values ( $\mathbf{v}^{\text{cmp}}$ ) required by NSA, we first pass the individual token keys  $\mathbf{k}_t$  and values  $\mathbf{v}_t$  within a block through a residual module (ResBlock), and then average the resulting features in-block (AvgPool):

$$\mathbf{k}^{\text{cmp}} = \text{AvgPool}(\text{ResBlock}(\mathbf{k}_t)), \quad \mathbf{v}^{\text{cmp}} = \text{AvgPool}(\text{ResBlock}(\mathbf{v}_t))$$

*Sparse Reconstruction Transformer* The sparse reconstruction transformer uses the same number of transformer blocks (24) and hidden dimensions (1024) as the Stage 1 model. The key difference is that we replace the standard CrossAttn in Eq. (3) and (4) with NSACrossAttn, which uses GPU-friendly Triton kernels



**Fig. 3: Attention-score vs. 3D-aware selection.** Standard attention captures 2D-3D correspondence only in deep layers, whereas our 3D-aware routing ensures stable selection to enhance texture.

to combine coarse global context with sparse fine-grained context through separate attention branches. The compressed KV-head layout is crucial for making these Triton kernels memory efficient; additional implementation details are included in the supplementary material. This structural alignment lets us initialize sparse-model weights directly from Stage 1, which significantly accelerates convergence. Furthermore, instead of predicting the high-resolution sparse volume from scratch, we train the sparse transformer to predict the residual over the dense prediction, an approach we found to be empirically much more effective. Let  $\mathbf{x}_d^{\text{up}}$  and  $\mathbf{y}_d^{\text{up}}$  denote the sparse tokens obtained by selecting and upsampling the informative regions of the dense predictions  $\mathbf{x}_d$  and  $\mathbf{y}_d$ . The forward pass through the sparse transformer blocks is defined as:

$$\begin{aligned} \mathbf{x}^{(m)}, \mathbf{y}^{(m)} = \text{NSABlock}^{(m)} & \left( \mathbf{x}^{(m-1)} + \text{Linear}^{(m)}(\mathbf{x}_d^{\text{up}}), \right. \\ & \left. \mathbf{y}^{(m-1)} + \text{Linear}^{(m)}(\mathbf{y}_d^{\text{up}}) \right), \quad \text{for } m = 1 \dots 24 \end{aligned} \quad (12)$$

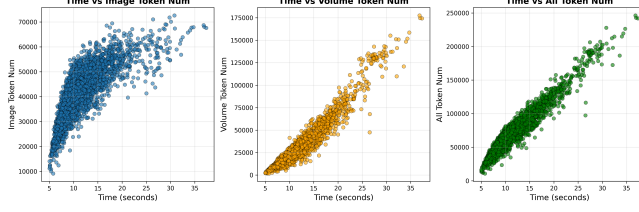
$$\mathbf{x}_s, \mathbf{y}_s = \mathbf{x}^{(24)} + \mathbf{x}_d^{\text{up}}, \mathbf{y}^{(24)} + \mathbf{y}_d^{\text{up}} \quad (13)$$

We project output tokens into a sparse feature volume  $\mathbf{X}_s$  using the same linear layer as Eq. (7), resulting in resolution  $S_f^{\text{vol}} = 4S^{\text{vol}} = 384$ . Note that  $\mathbf{X}_s$  only contains features computed from selected active volume tokens. We therefore combine it with the full dense feature volume  $\mathbf{X}_d$  during volume rendering. During ray marching, if a sampled point falls into a voxel  $(i, j, k)$  where the mask  $\mathbf{M}_{i,j,k}^{\text{vol}}$  is true, we query high-frequency features from  $\mathbf{X}_s$ . Conversely, if the point lies in empty space, we query the low-frequency features from  $\mathbf{X}_d$ . For points on the boundary between the two, we compute a linear combination of  $\mathbf{X}_s$  and  $\mathbf{X}_d$  to ensure smooth transitions. Finally, the lightweight MLP decoder processes these queried features exactly as in the dense reconstruction transformer.

### 3.3 3D-Aware Block Routing

Although our experiments show that LSRM benefits from an expanded context window and already improves fidelity over SOTA, standard NSA KV-block selection remains a bottleneck. In vanilla NSA, token blocks are chosen solely from attention scores against  $\mathbf{k}^{\text{cmp}}$ . As Fig. 3 illustrates, this data-driven routing often works in deeper layers, yet early transformer blocks frequently miss the most





**Fig. 4: Token count vs. training time.** Distribution of active tokens and their corresponding training times for resolutions  $S^{\text{img}} = 96$  and  $S^{\text{vol}} = 64$ . Each point represents a single data instance.

spatially relevant blocks, which cascades through the network and reduces final reconstruction quality.

To address this, we propose a 3D-aware routing strategy that selects KV-blocks using the coarse geometry established in Stage 1. We first assign an explicit 3D coordinate to each token. A volume token’s coordinate  $\mathbf{p}^{\text{vol}}$  is simply its voxel-center position. For an image token corresponding to an  $8 \times 8$  patch, its 3D location  $\mathbf{p}^{\text{img}}$  is defined as the foreground surface point with the highest opacity, rendered via the dense feature volume  $\mathbf{X}_d$ . We then define the 3D point sets  $\mathcal{P}^{\text{img}}$  and  $\mathcal{P}^{\text{vol}}$  for image and volume blocks as the collections of token coordinates in the block. Our geometry-based block-selection rules are:

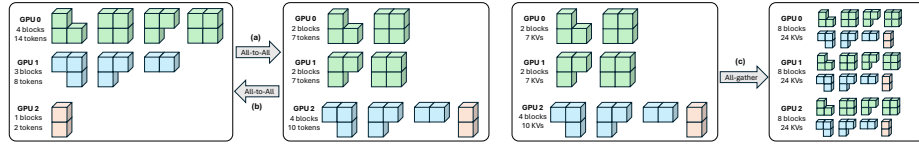
*Image/Volume to Image Blocks* For a given query token with 3D coordinate  $\mathbf{p} \in \{\mathbf{p}^{\text{img}}, \mathbf{p}^{\text{vol}}\}$ , we first project  $\mathbf{p}$  onto each input image plane using the known camera parameters. To limit computation, we retrieve the  $B^i$  image blocks per view whose 2D centers are closest to this projection. Next, for each candidate block we compute the minimal 3D Euclidean distance between  $\mathbf{p}$  and the point set  $\mathcal{P}^{\text{img}}$ . Finally, we select the  $B^{i2i}$  (image queries) or  $B^{v2i}$  (volume queries) candidates with the smallest such distances.

*Image/Volume to Volume Blocks* Given the query coordinate  $\mathbf{p} \in \{\mathbf{p}^{\text{img}}, \mathbf{p}^{\text{vol}}\}$ , we compute its 3D Euclidean distance to the spatial centers of all volume token blocks. We then select the  $B^{i2v}$  (for image queries) or  $B^{v2v}$  (for volume queries) volume blocks with the smallest center distances.

Fig. 3 compares our *volume-to-image* block selection against the original NSA mechanism on a converged model. The visualization confirms that our geometric approach establishes far more accurate 2D-3D correspondences. In Sec. 4, we show this enhanced spatially-aware routing translates to substantial improvements in reconstructed texture details (Fig. 7 and Tab. 1).

### 3.4 Custom Sequence Parallelism for Workload Balancing

LSRM’s sparse tokenization yields highly variable active-token counts across instances, producing straggler GPUs that stall training (Fig. 4) and tightly limit attainable resolution during multi-GPU training, since each step waits for the slowest worker. Early attempts to cap per-GPU tokens with reservoir sampling



**Fig. 5: Custom sequence parallelism.** Toy example assuming  $\leq 4$  tokens per spatial block across 3 GPUs. (a) Tokens are distributed evenly without breaking spatial blocks. (b) Tokens return to their source GPUs for decoding and volume rendering. (c) All-gather-KV ensures global context before every NSACrossAttn layer.

routinely dropped as many as one third of active tokens, which destabilized optimization. To balance this dynamic workload systematically, we adapt context parallelism [26] into a custom **block-aware sequence parallelism**. Unlike naive splits that break the block structure required by NSA, our scheme spreads tokens evenly across GPUs while strictly keeping each spatial block on a single device. This preserves block locality, so window attention and KV-block compression are computed entirely on-device, with zero cross-GPU communication.

To execute compressed (CmpAttn) and selected (SelAttn) attention over global context, tokens must still be routed efficiently. We use an All-gather-KV strategy [16, 45] that exploits our high KV compression. Concretely, we all-gather to replicate  $(\mathbf{k}^{\text{cmp}}, \mathbf{v}^{\text{cmp}})$  and  $(\mathbf{k}, \mathbf{v})$  on every GPU, while keeping queries  $(\mathbf{q})$  locally sharded. Since the KV cache is small, the communication cost of this gather is modest. The result is a favorable trade-off that lets us balance compute across 8 GPUs. As shown in Fig. 5, we perform an all-to-all collective before the first NSABlock to distribute token blocks globally, and then invoke all-gather on keys and values before SelAttn and CmpAttn.

### 3.5 Implementation, Training, and Inference Details

We train LSRM for novel-view synthesis (NVS) on 600K curated 3D models and then fine-tune it for inverse rendering. At each iteration, we randomly select 12–16 input views and 4 target views from 32 camera poses sampled on a sphere, following LIRM [38] and MeshLRM [58] to ensure robustness to diverse camera trajectories. NVS training follows three stages: (1) We train the dense reconstruction transformer ( $S_{\text{d}}^{\text{vol}} = 16, S_{\text{d}}^{\text{img}} = 32$ ), yielding a decoded volume resolution of  $S_{\text{df}}^{\text{vol}} = 64$  for  $256 \times 256$  inputs. (2) We freeze the dense model and use its predictions to train the sparse reconstruction transformer on the fly. We start at lower resolutions ( $S^{\text{img}} = 64, S^{\text{vol}} = 64$ ) to support larger batch sizes and quicker convergence. (3) We scale sparse resolutions to  $S^{\text{img}} = 96, S^{\text{vol}} = 96$ . This final stage relies on our custom block-aware sequence parallelism for stable training, with active image and volume token counts peaking at roughly 100K and 300K, respectively. Compared to recent SOTA like LIRM [38]—which processes a maximum of 8 images ( $64 \times 64 \times 8 \approx 32\text{K}$  image tokens) and uses a hexaplane representation ( $48 \times 48 \times 6 \approx 14\text{K}$  object tokens)—LSRM scales to  $> 2\times$  more image tokens and  $> 20\times$  more object tokens. These three stages



**Fig. 6: Qualitative NVS Results on the GSO Dataset.** LSRM significantly improves texture fidelity over prior SOTA [38]. Zoomed-in comparisons show that our method successfully recovers legible text (bottom left), complex geometry (top left) and sharp facial structures (right), whereas the baseline produces blurred artifacts.

require 5, 7, and 3 days, respectively, on 128 H100 GPUs. Across all stages, we optimize rendering, foreground mask, and depth losses, adding a numerical normal loss [38] during the final 500–1000 iterations to boost geometric fidelity.

Because our goal is to evaluate context-window scaling, we reduce compute by initializing inverse rendering from the converged NVS weights. We apply two task-specific architectural changes: (1) we encode background images with a linear layer [38] to better disentangle lighting from materials, and (2) we initialize the BRDF prediction head with pre-trained weights from  $\text{MLP}_c$ . We reuse the same 600K 3D models, rendering them under real HDR environment maps with extensive augmentations. Fine-tuning mirrors the three-stage NVS schedule. For efficiency, we initialize the first two stages directly from their NVS counterparts and halve the training iterations, while the final stage initializes from the converged second stage. More details are included in the supplementary materials.

All inference experiments are conducted on a single NVIDIA H100 GPU. With 16 input images at the highest resolution, LSRM requires 40 GB peak GPU memory, and the sparse reconstruction transformer reconstructs an object in about 5 seconds on average. This remains orders of magnitude faster than optimization-based alternatives [48, 52] and is comparable to LIRM’s iterative refinement when using the same number of views.

## 4 Experiments

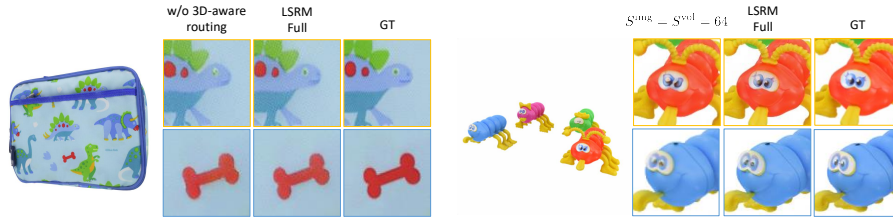
*Object-centric Novel View Synthesis* We evaluate the NVS quality of LSRM on the GSO dataset [14]. Inputs are 16 uniformly sampled  $768 \times 768$  views rendered with the camera configuration of [38], and we render 12  $768 \times 768$  novel views for evaluation. Following standard practice, ground-truth novel views are

**Table 1: Quantitative NVS Results on the GSO Dataset.** We compare LSRM against SOTA feed-forward baselines [38, 58, 73]. Our ablations evaluate the dense representation, context-window scaling through the image and volume token resolutions  $S^{\text{img}}$  and  $S^{\text{vol}}$ , and 3D-aware block routing. Best, second best, and third best results are highlighted in **red**, **orange**, and **yellow**, respectively.

| Metric                        | Baselines |       |       |       | Ours (LSRM Ablations) |                     |                                    |            |                                    |              |
|-------------------------------|-----------|-------|-------|-------|-----------------------|---------------------|------------------------------------|------------|------------------------------------|--------------|
|                               | Mesh      | GS    | LIRM  | Dense | $S^{\text{img}}=64$   | $S^{\text{img}}=48$ | $S^{\text{img}}=S^{\text{vol}}=64$ |            | $S^{\text{img}}=S^{\text{vol}}=96$ |              |
|                               | -LRM      | -LRM  |       | model | $S^{\text{vol}}=48$   | $S^{\text{vol}}=64$ | w/o routing                        | w/ routing | w/o routing                        | w/ routing   |
|                               |           |       |       |       |                       |                     |                                    |            |                                    |              |
| <b>PSNR</b> ( $\uparrow$ )    | 28.13     | 30.52 | 30.65 | 28.45 | 31.68                 | 31.82               | 31.34                              | 31.93      | 32.72                              | <b>33.08</b> |
| <b>SSIM</b> ( $\uparrow$ )    | 0.923     | 0.952 | 0.949 | 0.934 | 0.958                 | 0.959               | 0.956                              | 0.962      | 0.968                              | <b>0.971</b> |
| <b>LPIPS</b> ( $\downarrow$ ) | 0.093     | 0.050 | 0.054 | 0.079 | 0.044                 | 0.043               | 0.045                              | 0.040      | 0.032                              | <b>0.028</b> |

rendered at  $512 \times 512$ , and we downsample our predictions to this resolution before computing metrics. We compute PSNR, SSIM and LPIPS on the rendered RGB images, averaging over all objects and target views, and we use identical evaluation scripts across methods. Quantitative and qualitative results are summarized in Tab. 1 and Fig. 6, respectively. In Tab. 1, we compare LSRM against state-of-the-art feed-forward baselines. LIRM [38] uses the same number of input images via iterative refinement, but operates at a lower  $512 \times 512$  resolution. By contrast, LSRM increases the active image and volume token counts during reconstruction, yielding a substantial boost in overall fidelity. Our full model improves PSNR by  $> 2.4$  dB and reduces LPIPS by  $> 40\%$  relative to the strongest baseline. These gains are also evident qualitatively: LSRM produces sharper textures and recovers recognizable facial structure (e.g., the Little Mermaid and the bee). Even for objects with complex geometry (transformers, top left and bottom right), zoomed-in crops are nearly indistinguishable from ground truth. We additionally perform ablations to validate key components, including context-window scaling via  $S^{\text{img}}$  and  $S^{\text{vol}}$ , as well as our 3D-aware block routing. Results in Tab. 1 and Fig. 7 show that enlarging the context window provides the largest gains, while 3D-aware routing further sharpens fine-grained structures and restores previously distorted faces. Separately scaling the image-token and volume-token context lengths with 3D-aware routing shows that both contribute to reconstruction quality, with volume-token scaling having the larger impact: reducing  $S^{\text{vol}}$  from 64 to 48 drops PSNR to 31.68, compared to 31.82 when reducing  $S^{\text{img}}$  from 64 to 48.

*Object-centric Inverse Rendering* We evaluate LSRM’s inverse-rendering extension on three widely used benchmarks: StanfordORB [27], DigitalTwinCatalogue (DTC) [13], and ObjectsWithLighting (OWL) [53]. These datasets span diverse materials, illumination, and capture setups, offering a stringent test of relightability, texture fidelity, and geometric consistency. We focus our comparisons on optimization-based and feed-forward baselines with strong reported quantitative results or available qualitative comparisons. As summarized in Tab. 2, LSRM consistently outperforms prior feed-forward baselines. However, gains in pixel-wise metrics such as PSNR are comparatively modest. This reflects the inherently



**Fig. 7: Qualitative Ablation Study on the GSO Dataset.** Visual comparisons demonstrate that both our 3D-aware block routing and increased spatial resolutions independently enhance texture fidelity and overall rendering quality.

**Table 2: Quantitative Inverse Rendering Results.** We evaluate our approach against optimization-based and feed-forward baselines across three benchmarks. Notably, LSRM consistently achieves the lowest LPIPS across all datasets, a metric highly sensitive to fine texture quality and perceptual fidelity.

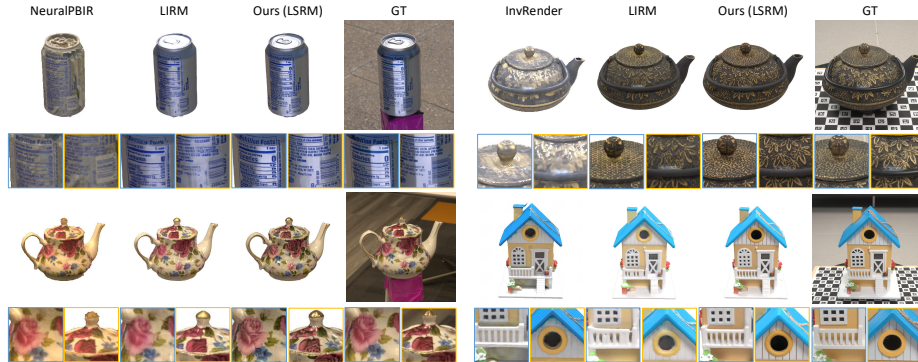
| Method             | StanfordORB [27] |              |              |              |             | DTC [13]     |              |              |              | OWL [53]     |             |              |  |
|--------------------|------------------|--------------|--------------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|-------------|--------------|--|
|                    | PSNR-H           | PSNR-L       | SSIM         | LPIPS        | CD          | PSNR-H       | PSNR-L       | SSIM         | LPIPS        | PSNR         | SSIM        | LPIPS        |  |
|                    | ↑                | ↑            | ↑            | ↓            | ↓           | ↑            | ↑            | ↑            | ↓            | ↑            | ↑           | ↓            |  |
| NVDiffrecMc        | 24.43            | 31.60        | 0.972        | 0.036        | 0.51        | 27.78        | 34.55        | 0.952        | 0.042        | 19.82        | 0.73        | 0.389        |  |
| InvRender          | 23.76            | 30.83        | 0.970        | 0.046        | 0.44        | 29.52        | <b>35.98</b> | 0.961        | 0.037        | 23.77        | 0.78        | 0.369        |  |
| NeuralPBIR         | <b>26.01</b>     | <b>33.26</b> | <b>0.979</b> | 0.023        | 0.43        | N/A          | N/A          | N/A          | N/A          | N/A          | N/A         | N/A          |  |
| LIRM               | 25.09            | 32.45        | 0.972        | 0.025        | 0.38        | 27.65        | 34.84        | 0.960        | 0.031        | 23.27        | 0.77        | 0.322        |  |
| <b>Ours (LSRM)</b> | 25.47            | 32.85        | 0.977        | <b>0.022</b> | <b>0.29</b> | <b>29.67</b> | 35.66        | <b>0.964</b> | <b>0.025</b> | <b>24.88</b> | <b>0.79</b> | <b>0.269</b> |  |

ill-posed nature of inverse rendering: although LSRM recovers high-fidelity geometry and texture details, global ambiguities—e.g., slight global shifts in albedo, roughness and metallicity maps—can disproportionately penalize PSNR. In contrast, on perceptual metrics like LPIPS, which emphasize high-frequency structure, LSRM shows substantial improvements, reaching performance comparable to SOTA optimization-based methods.

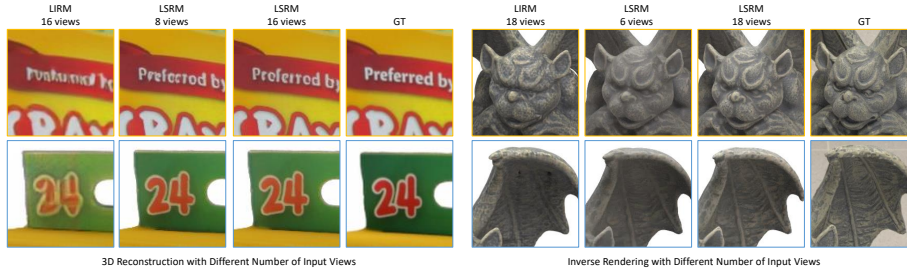
Qualitative results in Fig. 8 corroborate these gains in texture fidelity. For example, LSRM recovers legible text (the “Nutrition Facts” and “Calories 0” label on the can, top right) and fine textures (the golden leaves and flowers on the teapot, top right and bottom left). Beyond appearance, LSRM also improves geometry, accurately recovering the small fence in front of the birdhouse. This geometric accuracy is reflected in StanfordORB Chamfer Distance (CD), which drops markedly relative to LIRM and optimization-based approaches.

*Generalization to Smaller Number of Views* Since the primary focus of LSRM is to significantly scale the context window for both image and volume tokens, we maintain the number of input images between 12 and 16 during training and do not fine-tune our models on fewer images. Surprisingly, both our feed-forward 3D reconstruction and inverse rendering models generalize well to a smaller number of input views, outperforming the prior state-of-the-art, LIRM [38], which was explicitly trained on 4 to 8 images. We evaluate this generalization capability for novel-view synthesis on the GSO dataset [14]. As summarized in Tab. 3 and Fig.





**Fig. 8: Qualitative Inverse Rendering Results.** Visual comparisons against optimization-based and feed-forward baselines demonstrate that LSRM consistently recovers higher-fidelity textures and finer geometric details.



**Fig. 9: Qualitative Results with Fewer Input Views.** LSRM preserves high-fidelity geometry and texture when evaluated with reduced-view inputs for both novel-view synthesis and inverse rendering, even though it is trained with 12–16 views.

9, providing fewer views causes only a slight drop in texture sharpness compared to the full 16-view input. However, our results remain substantially superior to LIRM whether using 8 or 16 input views, as reflected in the quantitative metrics. Specifically, our 8-view results achieve close to a **2 dB** increase in PSNR compared to the 16-view LIRM, which corresponds to an approximate **37%** drop in MSE loss, while the LPIPS loss drops by more than **40%**. We observe a similar phenomenon in inverse rendering, where LSRM consistently outperforms LIRM using only 6 input views. Furthermore, evaluating on fewer views enables a fair comparison with RelitLRM [78], another feed-forward relighting model. Unlike our approach, RelitLRM does not explicitly predict material reflectance; rather, it relies on a generative network to synthesize novel lighting appearances, resulting in slower inference. Nevertheless, LSRM outperforms both RelitLRM and LIRM across two popular relighting benchmarks (see Tab. 4), with qualitative results in Fig. 9 corroborating these improvements.

| Metric                 | LIRM [38] |          | Ours (LSRM) |              |
|------------------------|-----------|----------|-------------|--------------|
|                        | 8 views   | 16 views | 8 views     | 16 views     |
| PSNR ( $\uparrow$ )    | 30.48     | 30.56    | 32.46       | <b>33.08</b> |
| SSIM ( $\uparrow$ )    | 0.947     | 0.948    | 0.968       | <b>0.971</b> |
| LPIPS ( $\downarrow$ ) | 0.056     | 0.054    | 0.031       | <b>0.028</b> |

**Table 3: Quantitative NVS Results with Fewer Input Views.** On GSO [14], LSRM maintains strong performance with 8 input views and substantially outperforms LIRM using either 8 or 16 views.

**Table 4: Quantitative Inverse Rendering Results with Fewer Input Views.** With only 6 input views, LSRM outperforms feed-forward relighting and inverse-rendering baselines across the evaluated benchmarks.

| Views | Method   | StanfordORB [27] |              |              |              |              | DTC [13]     |              |              |              | OWL [53]     |             |              |
|-------|----------|------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------|--------------|
|       |          | PSNR-H           | PSNR-L       | SSIM         | LPIPS        | CD           | PSNR-H       | PSNR-L       | SSIM         | LPIPS        | PSNR         | SSIM        | LPIPS        |
|       |          | $\uparrow$       | $\uparrow$   | $\uparrow$   | $\downarrow$ | $\downarrow$ | $\uparrow$   | $\uparrow$   | $\uparrow$   | $\downarrow$ | $\uparrow$   | $\uparrow$  | $\downarrow$ |
| 6     | RelitLRM | 24.67            | 31.52        | 0.969        | 0.032        | N/A          | N/A          | N/A          | N/A          | N/A          | 23.08        | 0.79        | 0.284        |
|       | LIRM     | 24.76            | 32.11        | 0.971        | 0.027        | 0.48         | 28.03        | 34.28        | 0.959        | 0.034        | 22.70        | 0.76        | 0.326        |
|       | Ours     | <b>25.57</b>     | <b>32.74</b> | <b>0.975</b> | <b>0.023</b> | <b>0.32</b>  | <b>29.09</b> | <b>35.16</b> | <b>0.961</b> | <b>0.028</b> | <b>24.43</b> | <b>0.80</b> | <b>0.255</b> |
| 18    | Ours     | 25.47            | 32.85        | 0.978        | 0.022        | 0.29         | 29.67        | 35.66        | 0.965        | 0.025        | 24.88        | 0.79        | 0.269        |

## 5 Conclusions

In this work, we show that scaling transformer context windows markedly improves feed-forward 3D reconstruction and inverse rendering. By scaling object and image tokens by  $20\times$  and  $2\times$  with NSA, we close much of the texture- and geometry-fidelity gap to optimization. To make this scale practical, we introduce 3D-aware block routing and block-aware sequence parallelism, boosting rendering quality while improving GPU utilization. Extensive experiments demonstrate that LSRM sets a new SOTA, recovering crisp text and faithful facial structures.

*Limitations and Future Work* LSRM introduces a general coarse-to-fine pipeline that leverages NSA to effectively scale context windows for 3D-related tasks. While this paper validates its effectiveness on object-centric 3D reconstruction and inverse rendering, we hope to extend this framework to large-scale scene reconstruction and general video/image generation in future work. Despite achieving SOTA reconstruction quality, several limitations remain. First, we observe that LSRM does not inherently resolve other ill-posed challenges in inverse rendering, such as accurately estimating roughness and metallic material properties. Furthermore, we still encounter cases where extremely fine details—such as small ingredient lists on cans—remain illegible. To address this, future iterations could expand the context window even further by integrating sequence parallelism techniques like Ulysses attention [22] and ring attention [45].

*Acknowledgements* We would like to thank David Clabaugh for creating the teaser figures, Dilin Wang and Yuchen Fan for their support with the training infrastructure, and Hao Tan, Sai Bi, and Kai Zhang for technical discussions regarding sparse attention.

## References

1. Ainslie, J., Lee-Thorp, J., De Jong, M., Zemlyanskiy, Y., Lebrón, F., Sanghai, S.: Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. pp. 4895–4901 (2023)
2. Bi, S., Xu, Z., Sunkavalli, K., Kriegman, D., Ramamoorthi, R.: Deep 3d capture: Geometry and reflectance from sparse multi-view images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5960–5969 (2020)
3. Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.: Nerd: Neural reflectance decomposition from image collections. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12684–12694 (2021)
4. Boss, M., Engelhardt, A., Kar, A., Li, Y., Sun, D., Barron, J., Lensch, H., Jampani, V.: Samurai: Shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems* **35**, 26389–26403 (2022)
5. Boss, M., Jampani, V., Braun, R., Liu, C., Barron, J., Lensch, H.: Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems* **34**, 10691–10704 (2021)
6. Careaga, C., Aksoy, Y.: Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics* **43**(1), 1–24 (2023)
7. Chen, A., Xu, H., Esposito, S., Tang, S., Geiger, A.: Lara: Efficient large-baseline radiance fields. In: *European conference on computer vision*. pp. 338–355. Springer (2024)
8. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624* (2025)
9. Chen, Z., Tang, J., Dong, Y., Cao, Z., Hong, F., Lan, Y., Wang, T., Xie, H., Wu, T., Saito, S., et al.: 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 26576–26586 (2025)
10. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. pp. 145–156 (2000)
11. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13142–13153 (2023)
12. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)* **37**(4), 1–15 (2018)
13. Dong, Z., Chen, K., Lv, Z., Yu, H.X., Zhang, Y., Zhang, C., Zhu, Y., Tian, S., Li, Z., Moffatt, G., et al.: Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 753–763 (2025)
14. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: *2022 International Conference on Robotics and Automation (ICRA)*. pp. 2553–2560. IEEE (2022)



15. Engelhardt, A., Raj, A., Boss, M., Zhang, Y., Kar, A., Li, Y., Sun, D., Brualla, R.M., Barron, J.T., Lensch, H., et al.: Shinobi: Shape and illumination using neural object decomposition via brdf optimization in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19636–19646 (2024)
16. Fang, J., Zhao, S.: Usp: A unified sequence parallelism approach for long context generative ai. arXiv preprint arXiv:2405.07719 (2024)
17. Goldman, D.B., Curless, B., Hertzmann, A., Seitz, S.M.: Shape and spatially-varying brdfs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(6), 1060–1071 (2009)
18. Gupta, A., Xiong, W., Nie, Y., Jones, I., Oğuz, B.: 3dgen: Triplane latent diffusion for textured mesh generation. arXiv preprint arXiv:2303.05371 (2023)
19. Hasselgren, J., Hofmann, N., Munkberg, J.: Shape, light, and material decomposition from images using monte carlo rendering and denoising. *Advances in Neural Information Processing Systems* **35**, 22856–22869 (2022)
20. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
21. Hunyuan3D, T., Yang, S., Yang, M., Feng, Y., Huang, X., Zhang, S., He, Z., Luo, D., Liu, H., Zhao, Y., et al.: Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. arXiv preprint arXiv:2506.15442 (2025)
22. Jacobs, S.A., Tanaka, M., Zhang, C., Zhang, M., Song, S.L., Rajbhandari, S., He, Y.: Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. arXiv preprint arXiv:2309.14509 (2023)
23. Jiang, Y., Tu, J., Liu, Y., Gao, X., Long, X., Wang, W., Ma, Y.: Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5322–5332 (2024)
24. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snively, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. arXiv preprint arXiv:2410.17242 (2024)
25. Jin, H., Liu, I., Xu, P., Zhang, X., Han, S., Bi, S., Zhou, X., Xu, Z., Su, H.: Tensor: Tensorial inverse rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2023)
26. Korthikanti, V.A., Casper, J., Lym, S., McAfee, L., Andersch, M., Shoeybi, M., Catanzaro, B.: Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems* **5**, 341–353 (2023)
27. Kuang, Z., Zhang, Y., Yu, H.X., Agarwala, S., Wu, E., Wu, J., et al.: Stanford-orb: a real-world 3d object inverse rendering benchmark. *Advances in Neural Information Processing Systems* **36** (2024)
28. Lan, Y., Hong, F., Zhou, S., Yang, S., Meng, X., Chen, Y., Lyu, Z., Dai, B., Pan, X., Loy, C.C.: Ln3diff++: Scalable latent neural fields diffusion for speedy 3d generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
29. Lan, Y., Zhou, S., Lyu, Z., Hong, F., Yang, S., Dai, B., Pan, X., Loy, C.C.: Gaussiananything: Interactive point cloud latent diffusion for 3d generation (2025)
30. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European conference on computer vision. pp. 71–91. Springer (2024)
31. Li, W., Liu, J., Yan, H., Chen, R., Liang, Y., Chen, X., Tan, P., Long, X.: Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. arXiv preprint arXiv:2405.14979 (2024)

32. Li, W., Zhang, X., Sun, Z., Qi, D., Li, H., Cheng, W., Cai, W., Wu, S., Liu, J., Wang, Z., et al.: Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. arXiv preprint arXiv:2505.07747 (2025)
33. Li, X., Dong, Y., Peers, P., Tong, X.: Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (ToG)* **36**(4), 1–11 (2017)
34. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., et al.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
35. Li, Z., Snavely, N.: Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 371–387 (2018)
36. Li, Z., Shi, J., Bi, S., Zhu, R., Sunkavalli, K., Hasan, M., Xu, Z., Ramamoorthi, R., Chandraker, M.: Physically-based editing of indoor scene lighting from a single image. In: *European Conference on Computer Vision*. pp. 555–572. Springer (2022)
37. Li, Z., Sunkavalli, K., Chandraker, M.: Materials for masses: Svbrdf acquisition with a single mobile phone image. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 72–87 (2018)
38. Li, Z., Wang, D., Chen, K., Lv, Z., Nguyen-Phuoc, T., Lee, M., Huang, J.B., Xiao, L., Zhu, Y., Marshall, C.S., et al.: Lirm: Large inverse rendering model for progressive reconstruction of shape, materials and view-dependent radiance fields. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 505–517 (2025)
39. Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)* **37**(6), 1–11 (2018)
40. Li, Z., Wang, Y., Zheng, H., Luo, Y., Wen, B.: Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. arXiv preprint arXiv:2505.14521 (2025)
41. Liang, H., Ren, J., Mirzaei, A., Torralba, A., Liu, Z., Gilitschenski, I., Fidler, S., Oztireli, C., Ling, H., Gojcic, Z., et al.: Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. arXiv preprint arXiv:2412.03526 (2024)
42. Liang, Z., Zhang, Q., Feng, Y., Shan, Y., Jia, K.: Gs-ir: 3d gaussian splatting for inverse rendering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21644–21653 (2024)
43. Lin, C.H., Lv, Z., Wu, S., Xu, Z., Nguyen-Phuoc, T., Tseng, H.Y., Straub, J., Khan, N., Xiao, L., Yang, M.H., et al.: Dgs-lrm: Real-time deformable 3d gaussian reconstruction from monocular videos. arXiv preprint arXiv:2506.09997 (2025)
44. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
45. Liu, H., Zaharia, M., Abbeel, P.: Ring attention with blockwise transformers for near-infinite context. arXiv preprint arXiv:2310.01889 (2023)
46. Ma, Z., Chen, X., Yu, S., Bi, S., Zhang, K., Ziwen, C., Xu, S., Yang, J., Xu, Z., Sunkavalli, K., et al.: 4d-lrm: Large space-time reconstruction model from and to any view at any time. arXiv preprint arXiv:2506.18890 (2025)
47. Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Seidel, H.P., Richardt, C., Theobalt, C.: Lime: Live intrinsic material estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 6315–6324 (2018)

48. Munkberg, J., Hasselgren, J., Shen, T., Gao, J., Chen, W., Evans, A., Müller, T., Fidler, S.: Extracting triangular 3d models, materials, and lighting from images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8280–8290 (2022)
49. Shi, Y., Wu, Y., Wu, C., Liu, X., Zhao, C., Feng, H., Liu, J., Zhang, L., Zhang, J., Zhou, B., et al.: Gir: 3d gaussian inverse rendering for relightable scene factorization. *arXiv preprint arXiv:2312.05133* (2023)
50. Siddiqui, Y., Monnier, T., Kokkinos, F., Kariya, M., Kleiman, Y., Garreau, E., Gafni, O., Neverova, N., Vedaldi, A., Shapovalov, R., et al.: Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *arXiv preprint arXiv:2407.02445* (2024)
51. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. *arXiv preprint arXiv:2508.10104* (2025)
52. Sun, C., Cai, G., Li, Z., Yan, K., Zhang, C., Marshall, C., Huang, J.B., Zhao, S., Dong, Z.: Neural-pbir reconstruction of shape, material, and illumination. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 18046–18056 (2023)
53. Ummenhofer, B., Agrawal, S., Sepulveda, R., Lao, Y., Zhang, K., Cheng, T., Richter, S., Wang, S., Ros, G.: Objects with lighting: A real-world dataset for evaluating reconstruction and rendering for object relighting. In: *2024 International Conference on 3D Vision (3DV)*. pp. 137–147. IEEE (2024)
54. Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., et al.: Lion: Latent point diffusion models for 3d shape generation. *Advances in neural information processing systems* **35**, 10021–10039 (2022)
55. Vaswani, A.: Attention is all you need. *Advances in Neural Information Processing Systems* (2017)
56. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5294–5306 (2025)
57. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 20697–20709 (2024)
58. Wei, X., Zhang, K., Bi, S., Tan, H., Luan, F., Deschaintre, V., Sunkavalli, K., Su, H., Xu, Z.: Meshlrn: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385* (2024)
59. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Xu, J., Torr, P., Cao, X., Yao, Y.: Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems* **37**, 121859–121881 (2024)
60. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Yang, Y., Bao, Y., Qian, J., Zhu, S., Cao, X., Torr, P., et al.: Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412* (2025)
61. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692* (2025)
62. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 21469–21480 (2025)

63. Xu, Y., Shi, Z., Yifan, W., Chen, H., Yang, C., Peng, S., Shen, Y., Wetzstein, G.: Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In: European Conference on Computer Vision. pp. 1–20. Springer (2024)
64. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R., Lv, Z.: 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. arXiv preprint arXiv:2506.08015 (2025)
65. Yang, H., Dong, Y., Jiang, H., Xu, D., Pavlakos, G., Huang, Q.: Atlas gaussians diffusion for 3d generation. arXiv preprint arXiv:2408.13055 (2024)
66. Yang, Z., Chen, Y., Gao, X., Yuan, Y., Wu, Y., Zhou, X., Jin, X.: Sire-ir: Inverse rendering for brdf reconstruction with shadow and illumination removal in high-illuminance scenes. arXiv preprint arXiv:2310.13030 (2023)
67. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems* **34**, 4805–4815 (2021)
68. Yuan, J., Gao, H., Dai, D., Luo, J., Zhao, L., Zhang, Z., Xie, Z., Wei, Y., Wang, L., Xiao, Z., et al.: Native sparse attention: Hardware-aligned and natively trainable sparse attention. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 23078–23097 (2025)
69. Zarzar, J., Monnier, T., Shapovalov, R., Vedaldi, A., Novotny, D.: Twinner: Shining light on digital twins in a few snaps. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5859–5869 (2025)
70. Zeng, Z., Deschaintre, V., Georgiev, I., Hold-Geoffroy, Y., Hu, Y., Luan, F., Yan, L.Q., Hašan, M.: Rgbx: Image decomposition and synthesis using material- and lighting-aware diffusion models. In: ACM SIGGRAPH 2024 Conference Papers. SIGGRAPH '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3641519.3657445>, <https://doi.org/10.1145/3641519.3657445>
71. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)* **42**(4), 1–16 (2023)
72. Zhang, C., Moing, G.L., Koppula, S., Rocco, I., Momeni, L., Xie, J., Sun, S., Sukthankar, R., Barral, J.K., Hadsell, R., et al.: Efficiently reconstructing dynamic scenes one d4rt at a time. arXiv preprint arXiv:2512.08924 (2025)
73. Zhang, K., Bi, S., Tan, H., Xiangli, Y., Zhao, N., Sunkavalli, K., Xu, Z.: Gs-lrm: Large reconstruction model for 3d gaussian splatting. In: European Conference on Computer Vision. pp. 1–19. Springer (2025)
74. Zhang, K., Luan, F., Li, Z., Snavely, N.: Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5565–5574 (2022)
75. Zhang, K., Luan, F., Wang, Q., Bala, K., Snavely, N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5453–5462 (2021)
76. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)* **43**(4), 1–20 (2024)
77. Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W.T., Tan, H.: Test-time training done right. arXiv preprint arXiv:2505.23884 (2025)
78. Zhang, T., Kuang, Z., Jin, H., Xu, Z., Bi, S., Tan, H., Zhang, H., Hu, Y., Hasan, M., Freeman, W.T., et al.: Relitlrm: Generative relightable radiance for large reconstruction models. arXiv preprint arXiv:2410.06231 (2024)

79. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)* **40**(6), 1–18 (2021)
80. Zhang, Y., Xu, T., Yu, J., Ye, Y., Jing, Y., Wang, J., Yu, J., Yang, W.: Nemf: Inverse volume rendering with neural microflake field. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 22919–22929 (2023)
81. Zhang, Y., Sun, J., He, X., Fu, H., Jia, R., Zhou, X.: Modeling indirect illumination for inverse rendering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18643–18652 (2022)
82. Zhao, Z., Liu, W., Chen, X., Zeng, X., Wang, R., Cheng, P., Fu, B., Chen, T., Yu, G., Gao, S.: Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in neural information processing systems* **36**, 73969–73982 (2023)
83. Ziwen, C., Tan, H., Zhang, K., Bi, S., Luan, F., Hong, Y., Fuxin, L., Xu, Z.: Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4349–4359 (2025)