NAMVIS: Next-Scale Autoregressive Multi-View Image Synthesis

Ramil Khafizov[1] **Ilya Statsenko**[2] **Ruslan Rakhimov**[2]
Artem Komarichev[1] **Peter Wonka**[3] **Evgeny Burnaev**[1,4]

[1]Applied AI Institute
[2]T-Tech
[3]KAUST
[4]AXXX

Abstract

Sparse-view novel view synthesis is a central problem in 3D content creation, but diffusion-based approaches remain limited by iterative denoising, making multi-view generation expensive at inference time. We introduce NAMVIS, a diffusion-free framework that reformulates multi-view image synthesis as geometry-conditioned next-scale autoregression. Instead of generating target views through repeated denoising, NAMVIS predicts discrete visual tokens through a small number of coarse-to-fine scale steps, while sampling all tokens within each scale and across target views in parallel. To anchor this generation process to explicit camera geometry, we propose Multi-scale Projective Pose Encoding, which injects source and target camera transformations into both target-view self-attention and source-to-target cross-attention at every resolution. NAMVIS further combines global conditioning with dense geometry-aware cross-attention, enabling the model to preserve source-view appearance while maintaining target-view consistency. Across Objaverse, GSO, and OmniObject3D, NAMVIS outperforms diffusion-based baselines in PSNR, SSIM, and LPIPS, while running over 3× faster than the evaluated diffusion baselines under the same evaluation setting. These results suggest that geometry-conditioned next-scale autoregression is a promising and efficient alternative to diffusion for sparse-view multi-view synthesis. Code, datasets, and additional results are available at `https://corl-team.github.io/namvis/`.

1 Introduction

Zero-shot novel view synthesis asks: given one or a few posed images of an unseen object, generate photorealistic images from new viewpoints—without per-scene optimization. The task is central to augmented reality, robotics, and 3D content creation [Poole et al., 2023, Tang et al., 2023], yet it is heavily under-constrained: a single photograph admits infinitely many consistent 3D scenes, so the model must supply strong priors over geometry, appearance, and illumination. In this work, we focus on posed sparse-view synthesis: given one or more source images with known cameras and a set of target cameras, the goal is to generate the corresponding target views without per-scene optimization or explicit 3D reconstruction.

Early methods reconstructed explicit 3D representations—NeRFs [Mildenhall et al., 2021], 3D Gaussian Splatting [Kerbl et al., 2023], or meshes [Yariv et al., 2021, Wang et al., 2021]—and required dozens to hundreds of views. These representations collapse in the few-image regime due to depth and shape ambiguities [Yu et al., 2021, Chan et al., 2022].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: **NAMVIS** performs sparse-view novel view synthesis from posed source images. Purple borders indicate input views; green borders indicate generated target views. NAMVIS generates target views with next-scale autoregressive decoding, producing sharp and more geometrically consistent results.

Zero-1-to-3 [Liu et al., 2023] showed that large pretrained 2D diffusion models [Rombach et al., 2022] already encode strong 3D priors: fine-tuning Stable Diffusion with relative pose conditioning enabled zero-shot NVS for the first time, spawning a wave of diffusion-based approaches [Shi et al., 2023, Qian et al., 2024, Long et al., 2024, Shi et al., 2024, Liu et al., 2024, Gu et al., 2023].

While diffusion-based methods have achieved strong visual quality, their iterative denoising process can make multi-view inference expensive, especially when generating several target views.

Meanwhile, next-scale autoregressive models have matched or surpassed diffusion quality in text-to-image generation [Tian et al., 2024, Han et al., 2025], yet their potential for 3D-aware synthesis remains largely untapped.

In this paper, we introduce *NAMVIS*, a purely autoregressive framework for zero-shot novel view synthesis from sparse observations. NAMVIS predicts target views through a denoising-free coarse-to-fine autoregression over discrete visual tokens, conditioned on a variable number of source images, their camera poses, and the desired target poses.

Our main contributions are:

- We introduce NAMVIS, a next-scale autoregressive framework for sparse-view multi-view synthesis that predicts target views over coarse-to-fine scale blocks, avoiding iterative denoising.
- We propose Multi-scale Projective Pose Encoding, which extends projective camera-based pose encoding to the varying spatial resolutions of next-scale autoregression and injects geometry into both self-attention and source-to-target cross-attention.
- We design a dual-path conditioning mechanism that combines global source-view conditioning with dense geometry-aware cross-attention to preserve both semantic appearance and fine spatial detail.
- We evaluate NAMVIS on Objaverse, GSO, and OmniObject3D against representative diffusion-based baselines, showing improved PSNR, SSIM, LPIPS, and faster inference under the same evaluation setting.
- To support reproducible sparse-view NVS research, we curate and release a large-scale rendered Objaverse-XL resource with RGBA images, camera poses, masks, point maps, generated captions, filtering metadata, and preprocessing scripts for approximately 200K objects.

2 Related work

Optimization-based and feed-forward novel view synthesis. A useful axis for organizing NVS research is the *generation paradigm*: methods that iterate (optimization-based or diffusion-based) versus methods that decode in a single pass (feed-forward or autoregressive). Classical NVS reconstructs an intermediate 3D representation—a NeRF [Mildenhall et al., 2021, Barron et al., 2022], voxel grid [Sitzmann et al., 2019], or mesh [Yariv et al., 2021, Wang et al., 2021]—and renders novel views from it. These methods excel when dozens of views are available but require per-scene optimization.

To bypass per-scene optimization, recent sparse-view reconstruction methods such as LRM [Hong et al., 2024] and LucidFusion [He et al., 2025] use feed-forward networks to predict explicit 3D representations, such as triplanes or 3D Gaussians, from sparse inputs. These methods address a closely related problem but differ in output space and inference objective: they first reconstruct an explicit 3D representation from which arbitrary views can be rendered, whereas NAMVIS directly synthesizes posed target images without producing a mesh, radiance field, or Gaussian representation. We therefore discuss feed-forward reconstruction methods as related work, while focusing our main experimental comparison on generative NVS methods that directly predict target views.

Diffusion-based generative NVS. A dominant generative NVS paradigm conditions pretrained 2D diffusion models on camera pose to implicitly capture 3D geometry. Zero-1-to-3 [Liu et al., 2023] pioneered this direction by fine-tuning Stable Diffusion [Rombach et al., 2022] with relative pose conditioning. Subsequent work improved multi-view consistency through cross-attention [Qian et al., 2024, Shi et al., 2023], synchronized multi-view generation [Liu et al., 2024], and joint prediction of multi-view appearance and geometry cues [Long et al., 2024]. More recently, EscherNet [Kong et al., 2024] introduced camera positional encodings to handle arbitrary numbers of reference and target views. Despite strong generation quality, diffusion-based NVS methods typically rely on iterative denoising at inference, which increases latency when generating multiple target views.

Autoregressive visual generation. Autoregressive (AR) models factorize image distributions as sequential token predictions. Early work [Van den Oord et al., 2016, Van Den Oord et al., 2017] operated in raster-scan order over discrete tokens. VAR [Tian et al., 2024] replaced token-by-token prediction with coarse-to-fine *next-scale* prediction, improving speed and quality. Infinity [Han et al., 2025] scaled this paradigm further with continuous-discrete bitwise representations. Recently, autoregressive formulations have been explored for multi-view generation. Notably, CausNVS [Kong et al., 2025] frames multi-view synthesis as an autoregressive sequence of *diffusion* steps, but still relies on iterative denoising for the generation of each individual view. NAMVIS bridges the gap between these domains by extending the *discrete* next-scale AR paradigm to 3D-aware multi-view generation. By injecting explicit 3D camera geometry via multi-scale ProPE and geometry-modulated cross-attention, NAMVIS provides a diffusion-free autoregressive alternative to diffusion-based sparse-view multi-view synthesis.

3 Method

We address zero-shot novel view synthesis. Given source RGB images I^{src} with camera poses $P^{\text{src}} \in SE(3)$ and target camera poses $P^{\text{tgt}} \in SE(3)$, we generate target images

$$\hat{I}^{\text{tgt}} = F_\theta(I^{\text{src}}, P^{\text{src}}, P^{\text{tgt}}).$$

NAMVIS instantiates F_θ as a next-scale autoregressive transformer built on Infinity [Han et al., 2025], which extends Visual Autoregressive (VAR) modeling [Tian et al., 2024]. We first review autoregressive next-scale models, then describe how we extend this design for pose-conditioned multi-view synthesis.

3.1 Background: Autoregressive Next-Scale Models

The core idea behind next-scale autoregression is to generate an image coarse-to-fine rather than pixel-by-pixel. A multi-scale VQ-VAE encodes an image I into a continuous feature map $f \in \mathbb{R}^{h \times w \times d}$ and K residual token maps $(r_1, \ldots, r_K)$, $r_k \in \mathbb{R}^{h_k \times w_k \times d}$, whose spatial resolution increases with k. In VAR [Tian et al., 2024], each residual is vector-quantized to a discrete codebook index; the

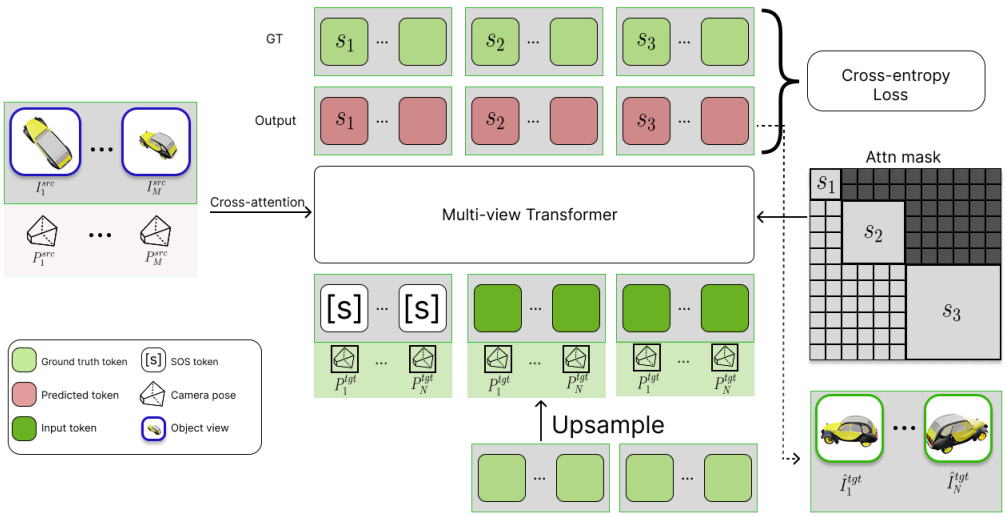

Figure 2: **NAMVIS architecture overview.** Source images are encoded by a frozen VQ-VAE into source features, which are pooled into a start-of-sequence token `[SOS]` and also serve as key–value pairs for dense cross-attention. The autoregressive transformer predicts target residual tokens scale by scale, with multi-scale ProPE injected into both self-attention (among target views) and cross-attention (from target to source). The frozen VQ-VAE decoder maps the final accumulated features to images.

continuous feature map is recovered by accumulating upsampled residuals:

$$f_k = \sum_{i=1}^{k} \mathrm{up}(r_i, (h, w)), \qquad f = f_K.$$

A transformer predicts these indices scale by scale, conditioned on an external signal c (e.g., a class label):

$$p(r_{1:K} \mid c) = \prod_{k=1}^{K} p(r_k \mid r_{<k}, c).$$

Infinity [Han et al., 2025] replaces codebook-based quantization with a bitwise representation: each token is quantized to a binary code $r_k(m, n) \in \{-1, +1\}^d$, creating an implicit vocabulary of size 2^d. The transformer predicts each bit independently via d binary classifiers, and a bitwise self-correction strategy stabilizes training.

3.2 NAMVIS

We show that multi-view image synthesis can be cast as geometry-conditioned next-scale prediction, eliminating the need for iterative denoising while preserving multi-view communication across target views. The key intuition is: if a coarse-to-fine autoregressive model can synthesize a photorealistic image from a text prompt, then we can replace the text conditioning with (i) visual features from reference views and (ii) 3D camera geometry, steering the same generative process to produce novel views. Figure 2 shows the overall pipeline; we describe each component below.

Training setup. During training, we sample random reference and target views from the same scene: source images I^{src} and target images I^{tgt} with camera poses P^{src} and P^{tgt}. A shared, frozen VQ-VAE encoder maps both into continuous feature maps $f^{\text{src}}, f^{\text{tgt}} \in \mathbb{R}^{h \times w \times d}$. We compute multi-scale residual token maps for the targets only: $(r_1^{\text{tgt}}, \ldots, r_K^{\text{tgt}})$, $r_k^{\text{tgt}} \in \mathbb{R}^{h_k \times w_k \times d}$, with the continuous feature at scale k reconstructed as $f_k^{\text{tgt}} = \sum_{i=1}^{k} \mathrm{up}(r_i^{\text{tgt}}, (h, w))$.

Autoregressive objective. The transformer predicts target residuals conditioned on source images and camera poses:

$$p_\theta(r^{\text{tgt}}_{1:K} \mid f^{\text{src}}, P^{\text{src}}, P^{\text{tgt}}) = \prod_{k=1}^{K} p_\theta\big(r^{\text{tgt}}_k \mid r^{\text{tgt}}_{<k}, f^{\text{src}}, P^{\text{src}}, P^{\text{tgt}}\big).$$

We retain Infinity's bitwise formulation and optimize a scale-reweighted cross-entropy loss over the independent bit predictions. At inference, NAMVIS performs a denoising-free coarse-to-fine decode over K scale steps: tokens within each scale and across target views are predicted in parallel, while later scales condition on earlier scales. To enable classifier-free guidance (CFG), we drop the source conditioning with probability 10% during training, replacing source features with a learned unconditional embedding.

Multi-view sequence layout and masking. To generate N target views concurrently, we interleave token sequences scale by scale. The combined sequence is $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_K)$, where $\mathcal{S}_k$ contains the flattened tokens of all N views at scale k. A block-causal attention mask enforces the autoregressive property: tokens in $\mathcal{S}_k$ attend to all tokens in $\mathcal{S}_{\leq k}$ (across all views) but are masked from $\mathcal{S}_{>k}$. Within a scale block, tokens from different views attend to each other freely, enabling multi-view communication at every resolution. A learned scale-level embedding $e_k \in \mathbb{R}^d$ is added before the first transformer block so that the model can distinguish the current resolution.

Multi-scale Projective Pose Encoding. Standard 2D RoPE captures image-plane position but does not encode the 3D relationship between cameras. We therefore use a multi-scale projective positional encoding based on ProPE [Li et al., 2025], adapted to next-scale autoregressive multi-view generation.

For a token feature x_s at autoregressive scale s, we split each attention head into one projective part and two local rotary parts:

$$x_s = [x_s^{\text{proj}},\ x_s^u,\ x_s^v].$$

For view v and scale s, our projective-RoPE transform is

$$\Psi_b^{v,s}(x_s) = \left[M_b(P_v)x_s^{\text{proj}},\ \text{RoPE}(x_s^u, u_s),\ \text{RoPE}(x_s^v, v_s)\right],$$

where (u_s, v_s) are the token-grid coordinates at scale s, P_v is the projective camera matrix for view v, and $b \in \{q, kv, o\}$ denotes the attention branch. Following ProPE, the branch-specific projective matrices are

$$M_q(P_v) = P_v^\top, \qquad M_{kv}(P_v) = P_v^{-1}, \qquad M_o(P_v) = P_v.$$

Thus, the projective component injects 3D camera geometry, while the two RoPE components preserve local 2D layout at the current autoregressive resolution. Because NAMVIS generates tokens over multiple scales, we construct separate coordinate grids and rotary coefficients for each target scale s.

We then use the same projective-RoPE operator in two attention routes. For target self-attention, queries, keys, values, and outputs all use the target-view camera:

$$Y^{\text{self}}_{t,s} = \Psi_o^{t,s}\left(\text{Attn}\left(\Psi_q^{t,s}(Q_{t,s}), \Psi_{kv}^{t,\leq s}(K_{t,\leq s}), \Psi_{kv}^{t,\leq s}(V_{t,\leq s})\right)\right).$$

The block-causal next-scale mask is applied inside this attention: tokens at scale s may attend to previous scales and to all target views at the current scale, but not to future scales.

For source-to-target cross-attention, target tokens are queries and source features are keys and values. Therefore the query and output branches use the target camera, while the key/value branches use the source cameras:

$$Y^{\text{cross}}_{t,s} = \Psi_o^{t,s}\left(\text{Attn}\left(\Psi_q^{t,s}(Q_{t,s}), \Psi_{kv}^{src}(K_{src}), \Psi_{kv}^{src}(V_{src})\right)\right).$$

Here, source features remain at the dense VQ-VAE feature resolution rather than being downsampled to the current target scale. This lets every target scale attend to high-resolution source evidence while preserving the coarse-to-fine autoregressive structure.

This role-aware routing makes geometry available in both forms of communication: target self-attention encourages consistency among generated views, while source-to-target cross-attention anchors generation to posed source observations. In target self-attention, every token belongs to a generated target view.

Image conditioning. Reference images enter the transformer through two complementary paths—global and dense—serving low-frequency semantic alignment and high-frequency spatial detail respectively.

Global path. A learnable query $q \in \mathbb{R}^d$ cross-attends to the source features f^{src}, producing a pooled vector:

$$\texttt{[SOS]} = \text{CrossAttn}(q, f^{\text{src}}).$$

This vector is prepended to the target sequence as a start-of-sequence token and is also fed through a mapping network to produce scale and shift parameters for Adaptive Layer Normalization (AdaLN), modulating every transformer block. When generating multiple target views, we duplicate the `[SOS]` token for each output sequence.

Dense path. A single pooled token cannot capture high-frequency detail, so every transformer block also performs dense cross-attention against the unpooled VQ-VAE features f^{src}. This cross-attention is modulated by ProPE (as described above), so the transformer aggregates source features according to the geometric relationship between source and target views. Crucially, while the target queries are processed in a progressive, multi-scale fashion, the source keys and values are maintained at their full, single-scale resolution. This asymmetry ensures that even the coarsest stages of autoregressive generation are strictly anchored by high-fidelity spatial details from the source views.

4 Experiments

4.1 Implementation Details

Datasets. We train our model on a curated subset of Objaverse-XL [Deitke et al., 2023a]. Since raw Objaverse-XL contains many assets with rendering failures, missing textures, low foreground coverage, and degenerate geometry, we apply a three-stage filtering pipeline. First, we remove assets with failed renders, low foreground occupancy, missing or predominantly white textures, and severe geometry artifacts. Second, we score rendered views using an aesthetic-quality model and discard assets with consistently low visual quality. Third, we generate image captions for the remaining assets and remove samples with degenerate captions, such as empty, generic, or visually inconsistent descriptions. This combines geometry-, visibility-, appearance-, and caption-based filtering.

The resulting dataset contains approximately 200K objects. For each object, we render 100 posed RGBA views and store the corresponding camera intrinsics, camera extrinsics, masks, point maps, and generated captions. To support reproducibility and future research, we publicly release the filtered scene list, rendered RGBA images, camera poses, masks, point maps, captions, and preprocessing scripts.

Architecture. NAMVIS is built upon the Infinity transformer architecture. The underlying multi-scale VQ-VAE is pre-trained and kept frozen during transformer training.

To reduce inference cost, we use standard KV caching together with a global ProPE caching mechanism. Since NAMVIS injects multi-scale projective camera encodings into multiple attention blocks, recomputing the geometry-dependent ProPE transformations at every layer would introduce redundant overhead. We therefore compute the ProPE projection tensors once for each source–target camera configuration and autoregressive scale, cache them globally, and reuse them across transformer layers. This reduces the cost of geometry-aware attention without changing the model outputs.

Training details. We train only the autoregressive transformer while keeping the VQ-VAE frozen. The model is optimized with AdamW using a base learning rate of 1×10^{-4}, a global batch size of 512, and 20 warm-up epochs, followed by cosine learning-rate decay. Training is run for 500 epochs in total. During training, we drop source-view conditioning with probability 0.1 to enable classifier-free guidance at inference time.

Evaluation metrics. We evaluate novel-view fidelity using PSNR and SSIM, where higher is better, and perceptual similarity using LPIPS, where lower is better. All metrics are computed between generated target views and the corresponding ground-truth renders at 256×256 resolution. To assess multi-view consistency, we additionally report COLMAP reconstructability, measured by the average number of reconstructed sparse points from the generated multi-view outputs.

Table 1: Quantitative comparison across Objaverse, GSO, and OmniObject3D.

Method	Objaverse (In-Domain)			GSO (Zero-Shot)			OmniObject3D (OO3D)			Time (s) ↓
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
Zero-1-to-3 [Liu et al., 2023]	19.097	0.816	0.152	15.614	0.749	0.257	16.094	0.768	0.219	2.6
Zero-1-to-3 XL [Deitke et al., 2023a]	19.341	0.827	0.136	15.936	0.747	0.236	16.320	0.767	0.209	2.6
Wonder3D [Long et al., 2024]	15.801	0.821	0.209	14.158	0.777	0.285	14.290	0.786	0.243	2.0
SyncDreamer [Liu et al., 2024]	19.368	0.826	0.156	16.961	0.781	0.246	16.766	0.793	0.207	16.6
EscherNet [Kong et al., 2024]	19.078	0.816	0.147	18.878	0.799	0.156	17.767	0.780	0.167	2.1
NAMVIS 1B (Ours)	**22.485**	**0.861**	**0.091**	**21.715**	**0.843**	**0.111**	**21.098**	**0.845**	**0.104**	**0.6**

4.2 Baselines

To evaluate the zero-shot novel view synthesis capabilities of NAMVIS, we compare against representative public diffusion-based NVS baselines that directly generate target views from posed source images. Zero-1-to-3 [Liu et al., 2023] and its scaled-up version Zero-1-to-3 XL [Deitke et al., 2023a] serve as foundational single-view baselines that synthesize a novel view using relative camera conditioning. We also compare against recent multi-view diffusion methods: Wonder3D [Long et al., 2024], which predicts multi-view appearance and geometry cues; SyncDreamer [Liu et al., 2024], which synchronizes multi-view generation to improve consistency; and EscherNet [Kong et al., 2024], which uses camera positional encoding (CaPE) to support flexible target-view generation. We use the official pretrained weights for all baselines.

For single-view baselines, we evaluate each available source view independently and report the best result according to PSNR. Some baselines impose restrictions on the camera elevation. Wonder3D [Long et al., 2024] requires source views with 0° elevation, whereas SyncDreamer [Liu et al., 2024] uses views at 30° elevation. Therefore, we evaluate methods under their supported camera protocols: Wonder3D is evaluated at 0° elevation, while the remaining methods are evaluated at 30° elevation.

4.3 Quantitative Results

We evaluate our method on Objaverse [Deitke et al., 2023b], GSO [Downs et al., 2022], and OmniObject3D (OO3D) [Wu et al., 2023] at a resolution of 256×256. Quantitative results are summarized in Table 1. We report metrics averaged across nine source–target view configurations: 1-to-1, 1-to-2, 1-to-3, 2-to-1, 2-to-2, 2-to-3, 3-to-1, 3-to-2, and 3-to-3. For each object, source and target views are sampled from a fixed set of available posed views. Full per-configuration results are provided in Appendix B.

The exact evaluation scenes for OO3D, GSO, and Objaverse are listed in Appendix H.

Perceptual Quality. NAMVIS consistently outperforms diffusion-based baselines across all three datasets. On the in-domain Objaverse benchmark, NAMVIS achieves 22.485 PSNR, 0.861 SSIM, and 0.091 LPIPS, improving over Zero-1-to-3, Zero-1-to-3 XL, Wonder3D, SyncDreamer, and EscherNet. The gains are especially pronounced in LPIPS, indicating that NAMVIS produces perceptually sharper and more faithful novel views. On the zero-shot GSO benchmark, NAMVIS reaches 21.715 PSNR, 0.843 SSIM, and 0.111 LPIPS, suggesting strong generalization beyond the Objaverse training distribution. NAMVIS also performs best on OmniObject3D, achieving 21.098 PSNR, 0.845 SSIM, and 0.104 LPIPS, suggesting that the learned autoregressive multi-view prior transfers to more realistic object scans.

Inference Efficiency. In addition to improving image-quality metrics, NAMVIS substantially reduces inference latency. In the 1-to-1 setting, NAMVIS requires 0.6 seconds per target view, while the fastest evaluated diffusion baseline, Wonder3D, requires 2.0 seconds. Thus, NAMVIS is approximately $3.3\times$ faster than the fastest diffusion baseline under the same 256×256 evaluation resolution.

View-count scaling. We further evaluate NAMVIS with up to six source and six target views on Objaverse in Table 2. NAMVIS supports $6\rightarrow6$ joint generation while achieving 26.12 PSNR, 0.9015 SSIM, and 0.0473 LPIPS, with 1.56 s inference time and 5.0 GB peak VRAM. This indicates that the proposed multi-view autoregressive formulation scales to larger source and target sets without excessive inference cost.

Table 2: Scaling to larger source and target view counts on Objaverse.

Setup	LPIPS ↓	SSIM ↑	PSNR ↑	Time (s) ↓	VRAM (GB) ↓
2→6	0.0654	0.8845	24.49	1.53	4.3
6→2	0.0457	0.9017	26.07	0.84	4.1
6→6	0.0473	0.9015	26.12	1.56	5.0

Cross-dataset generalization. All methods degrade when moving from Objaverse to GSO and OmniObject3D, but NAMVIS degrades more gracefully than the evaluated baselines. This suggests that geometry-conditioned next-scale autoregression learns a transferable object-level multi-view prior under our evaluation protocol.

Multi-View Consistency and Reconstructability. Standard 2D metrics such as PSNR, SSIM, and LPIPS evaluate per-view fidelity, but often overlook cross-view inconsistencies. To address this, we follow SyncDreamer [Liu et al., 2024] and evaluate geometric consistency through COLMAP-based reconstructability. We run COLMAP on the source and generate target views for all methods using identical settings. This metric serves as a proxy for multi-view geometric consistency, as inconsistent views hinder stable feature tracking, camera registration, and triangulation.

As shown in Table 3, NAMVIS reconstructs an average of 221 sparse points (counting failed reconstructions as zero), outperforming EscherNet (214), Wonder3D (172), and SyncDreamer (151). These results indicate that our jointly generated views support more stable feature matching and triangulation than existing multi-view diffusion baselines. We exclude Zero-1-to-3 [Liu et al., 2023] and Zero-1-to-3 XL [Deitke et al., 2023a] from this comparison, as they do not natively generate jointly conditioned multi-view sets.

Table 3: COLMAP reconstructability comparison. Higher is better.

Method	COLMAP Points ↑
Wonder3D [Long et al., 2024]	172
SyncDreamer [Liu et al., 2024]	151
EscherNet [Kong et al., 2024]	214
NAMVIS 1B (Ours)	**221**

Registration coverage We additionally measure camera-registration success across 30 scenes using identical COLMAP settings and report results in Table 4. NAMVIS registers 162 of 240 views (67.5%), compared with 145 of 240 (60.4%) for EscherNet, and approaches the 70.0% coverage obtained using ground-truth images.

Table 4: COLMAP camera-registration coverage across 30 scenes.

Method	Registered views	Coverage ↑
GT	168 / 240	70.0%
EscherNet	145 / 240	60.4%
NAMVIS	162 / 240	**67.5%**

In addition, we provide qualitative dense camera-trajectory videos in the supplement, where NAMVIS produces stable geometry and appearance across smooth viewpoint changes.

4.4 Qualitative Results

Figures 3 and 4 show qualitative comparisons on OO3D [Wu et al., 2023] and GSO [Downs et al., 2022]. NAMVIS better preserves the source-view appearance while producing target views that follow the requested camera poses. To accommodate baseline constraints, Wonder3D [Long et al., 2024] is evaluated at 0° camera elevation and all other baselines at 30°, while our method is evaluated at both angles. Compared with diffusion-based baselines, NAMVIS produces sharper object boundaries and more consistent geometry, especially for texture-rich and thin-structure objects.

Additional qualitative generations, source-view-count examples, and attention-map visualizations are provided in Appendices F, C, and D.

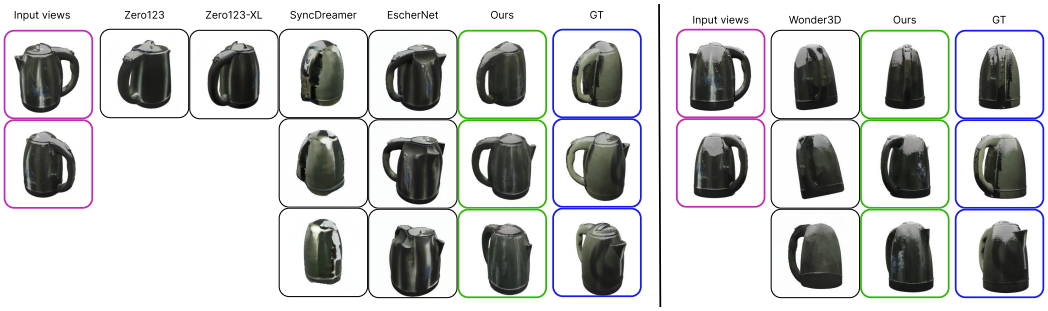

Figure 3: *Qualitative comparison on OO3D [Wu et al., 2023].*

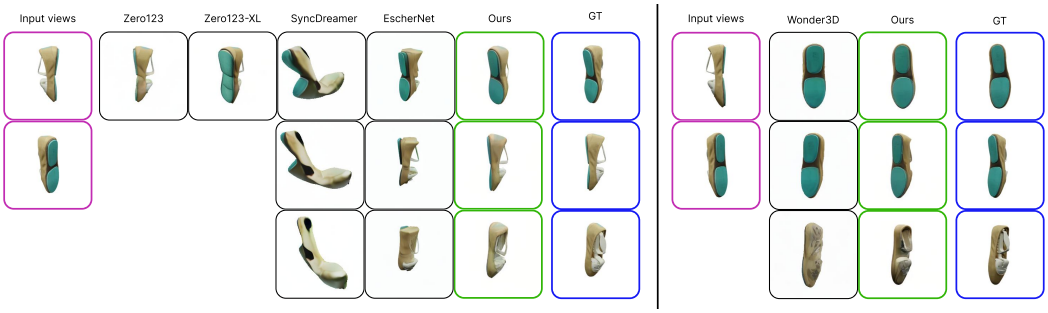

Figure 4: *Qualitative comparison on GSO [Downs et al., 2022].*

Ablations. We conduct ablations on a 10K-object subset of our curated Objaverse-XL data. We ablate full-parameter fine-tuning as shown in Table 5 by replacing it with LoRA [Hu et al., 2022] with rank $r = 256$. LoRA tuning performs substantially worse than full fine-tuning, suggesting that adapting a pretrained autoregressive image prior to geometry-conditioned multi-view synthesis requires updating the core transformer weights rather than only learning low-rank residual updates.

Additional results provided in Appendix E show that ProPE outperforms Plücker ray conditioning, dense cross-attention is critical for preserving fine spatial detail, and VQ-VAE source features outperform CLIP or separately learned dense features.

4.5 Limitations and future work

NAMVIS is trained and evaluated at 256 × 256, primarily due to our available compute budget. While the architecture can in principle support higher resolutions, increasing the resolution to 512 × 512 would increase the token count and training cost. Moreover, since the current checkpoint was trained only at 256 × 256, reliable higher-resolution generation would require additional training or fine-tuning.

A promising direction is to combine NAMVIS with more efficient scale-wise transformer designs. Recent work such as Switti [Voronov et al., 2025] suggests that next-scale generation may not require full causal attention to all previous scales, since the current-scale representation already contains accumulated lower-scale information. In NAMVIS, an analogous design would replace full block-causal attention over all previous target scales with scale-local or windowed target-target attention, while preserving geometry-aware source-to-target cross-attention. This could reduce memory and KV-cache cost at high resolutions without changing the core geometry-conditioned coarse-to-fine formulation.

Separately, our current model is trained on object-centric Objaverse-XL data, so its learned geometric priors may not transfer directly to complex, real-world scenes. However, the NAMVIS architecture itself is not restricted to object-centric inputs. Addressing this domain gap via larger, more diverse multi-view datasets that combine object-centric and scene-centric data is an important direction for future research.

Table 5: *Ablation study.* Each row replaces or removes one component of the full NAMVIS architecture. All variants are trained on a 10K subset of Objaverse-XL for 200 epochs.

Model Variant	PSNR ↑	SSIM ↑	LPIPS ↓
Geometry & Conditioning			
w/ Plücker Rays (instead of ProPE)	20.260	0.899	0.142
Prefix Conditioning (instead of Cross-Attn)	14.539	0.797	0.427
Feature Aggregation			
CLIP `[SOS]` (instead of VQ-VAE)	20.613	0.901	0.126
Learned Dense Features (instead of VQ-VAE)	20.886	0.901	0.134
w/o Dense Cross-Attn (`[SOS]` only)	17.891	0.865	0.248
Optimization			
LoRA Tuning (instead of Full Tuning)	18.912	0.883	0.185
Full NAMVIS (Ours)	**21.360**	**0.907**	**0.115**

5 Conclusion

We presented NAMVIS, a geometry-conditioned next-scale autoregressive framework for zero-shot novel view synthesis from sparse posed observations. Instead of generating target views through iterative diffusion denoising, NAMVIS performs coarse-to-fine autoregressive decoding over visual tokens, predicting all tokens within each scale and across target views in parallel. This provides an efficient diffusion-free alternative for sparse-view multi-view synthesis.

Across Objaverse, GSO, and OmniObject3D, NAMVIS improves PSNR, SSIM, and LPIPS over representative public diffusion-based baselines, including both single-view and multi-view generation methods. In the 1-to-1 setting, NAMVIS runs in 0.6 s per target view, approximately $3.3\times$ faster than the fastest evaluated diffusion baseline under the same 256×256 resolution setting. Together with the ablation studies and COLMAP reconstructability evaluation, these results suggest that geometry-conditioned next-scale autoregression is a promising direction for efficient sparse-view novel view synthesis.

Future work will focus on reducing sequence-length bottlenecks for higher resolutions and denser target-view generation, improving robustness to pose noise, and extending training beyond object-centric data to more complex scene-level multi-view settings.

Acknowledgments

This work was supported by a grant for research centers in the field of artificial intelligence provided by the Ministry of Economic Development of the Russian Federation under Agreements No. 000000C313925P4F0002 and No. 139-10-2025-033.

Competing Interests. The authors declare no competing interests.

References

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. In *Advances in Neural Information Processing Systems*, volume 36, pages 35799–35813, 2023a.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023b.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google Scanned Objects: A high-quality dataset of 3D scanned household items. In *International Conference on Robotics and Automation (ICRA)*, pages 2553–2560, 2022.

Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. NerfDiff: Single-image view synthesis with NeRF-guided distillation from 3D-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise AutoRegressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15733–15744, 2025.

Hao He, Yixun Liang, Luozhou Wang, Yuanhao Cai, Xinli Xu, Hao-Xiang Guo, Xiang Wen, and Yingcong Chen. LucidFusion: Reconstructing 3D Gaussians with arbitrary unposed images. *Computer Graphics Forum*, 44(7):e70227, 2025.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *International Conference on Learning Representations*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024.

Xin Kong, Daniel Watson, Yannick Strümpler, Michael Niemeyer, and Federico Tombari. CausNVS: Autoregressive multi-view diffusion for flexible 3D novel view synthesis. *arXiv preprint arXiv:2509.06579*, 2025.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems*, volume 38, pages 18020–18045, 2025.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations*, 2024.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11966–11976, 2022.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024.

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. *Communications of the ACM*, 65(1):99–106, 2021.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations*, 2023.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *International Conference on Learning Representations*, 2024.

Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. DeepVoxels: Learning persistent 3D feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-It-3D: High-fidelity 3D creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22819–22829, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, volume 37, pages 84839–84865, 2024.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with PixelCNN decoders. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2025.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems*, volume 34, pages 4805–4815, 2021.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations*, 2024.

A Architectural Details

NAMVIS uses a 1B-parameter transformer initialized from the Infinity [Han et al., 2025] 2B checkpoint. To obtain the 1B model, we keep every other transformer block from the 2B model and discard the remaining blocks. Each retained transformer block contains self-attention followed by source-to-target cross-attention and an MLP. The final model has 16 transformer blocks, hidden dimension 2048, and 16 attention heads.

B Runtime and View-Count Scaling

We evaluate NAMVIS for different numbers of source and target views. Table 6 reports inference time and image-quality metrics for each source–target configuration, where setup a–b denotes a source views and b target views. Inference time is measured at 256×256 resolution on a single NVIDIA A40 GPU.

Increasing the number of source views generally improves reconstruction quality, as reflected by lower LPIPS and higher SSIM and PSNR scores. This is expected, since additional source views reduce ambiguity in unobserved regions. Increasing the number of target views primarily affects computational cost: runtime increases as more target views are generated jointly, while image quality remains relatively stable for a fixed number of source views.

Overall, these results show that NAMVIS scales efficiently with the number of target views while benefiting from additional source-view evidence.

Caching breakdown. Table 7 reports the effect of caching on inference time in the 3-to-3 setting. Standard KV caching reduces repeated attention computation across autoregressive scales, while ProPE caching avoids recomputing geometry-dependent projective transformations across transformer layers. Removing KV caching increases runtime from 0.96 s to 1.06 s, while removing ProPE caching increases runtime to 1.28 s, indicating that caching geometry-dependent ProPE features is particularly important for efficient multi-view inference.

C Effect of the Number of Source Views

Figures 5 and 6 show how increasing the number of source views affects generation quality. As additional views become available, NAMVIS shifts from hallucinating unobserved regions to rendering object parts that are directly supported by the input views.

D Attention Maps Analysis

Visualizations of self-attention and cross-attention maps suggest that NAMVIS learns view-correspondence patterns consistent with the underlying camera geometry. For a query token in a generated target image, the model often assigns high attention weights to visually and geometrically corresponding regions across other views. We also observe a scale-dependent pattern: self-attention tends to emphasize target-target consistency at finer scales (Figure 7), while cross-attention focuses on source-target alignment at coarser scales (Figure 8).

Table 6: Inference time and image quality for different source–target view configurations. Setup a–b denotes a source views and b target views.

Dataset	Setup	NAMVIS				EscherNet [Kong et al., 2024]			
		Time (s)	LPIPS↓	SSIM↑	PSNR↑	Time (s)	LPIPS↓	SSIM↑	PSNR↑
GSO	1–1	0.60	0.1634	0.8104	19.0542	2.09	0.2924	0.7078	14.0517
	1–2	0.81	0.1668	0.8100	19.2163	2.16	0.2705	0.7334	14.5864
	1–3	0.94	0.1698	0.8110	19.1860	2.35	0.2683	0.7342	14.8349
	1–4	1.11	0.1633	0.8147	19.5188	2.78	0.2498	0.7498	15.6767
	2–1	0.62	0.0810	0.8603	23.0655	2.08	0.0927	0.8360	21.2183
	2–2	0.79	0.0883	0.8538	22.4639	2.17	0.1037	0.8336	20.6660
	2–3	0.95	0.1048	0.8451	21.8491	2.44	0.1208	0.8200	19.8665
	2–4	1.11	0.1061	0.8452	21.9103	2.81	0.1182	0.8233	20.1216
	3–1	0.63	0.0737	0.8675	23.5606	2.20	0.0890	0.8366	21.2918
	3–2	0.80	0.0725	0.8692	23.6664	2.23	0.0808	0.8488	21.8496
	3–3	0.96	0.0810	0.8632	23.3698	2.47	0.0890	0.8415	21.5421
	3–4	1.12	0.0835	0.8608	23.3741	2.84	0.0936	0.8393	21.5674
	4–1	0.63	0.0721	0.8704	23.8080	2.25	0.0846	0.8382	21.5261
	4–2	0.82	0.0691	0.8712	23.9583	2.27	0.0778	0.8501	22.1110
	4–3	0.96	0.0678	0.8708	24.1201	2.52	0.0737	0.8511	22.3149
	4–4	1.12	0.0666	0.8726	24.3523	2.98	0.0722	0.8532	22.5996
OO3D	1–1	0.60	0.1505	0.8095	18.6618	2.09	0.2842	0.7151	13.6204
	1–2	0.81	0.1529	0.8098	18.4677	2.16	0.2772	0.7345	14.1137
	1–3	0.94	0.1644	0.8021	18.0196	2.35	0.2815	0.7367	13.9419
	1–4	1.11	0.1545	0.8091	18.4509	2.78	0.2627	0.7455	14.4499
	2–1	0.62	0.0791	0.8609	22.2238	2.08	0.1254	0.8208	19.1116
	2–2	0.79	0.0851	0.8598	22.0147	2.17	0.1146	0.8353	19.3928
	2–3	0.95	0.0949	0.8513	21.3968	2.44	0.1244	0.8267	18.9607
	2–4	1.11	0.0907	0.8525	21.5572	2.81	0.1223	0.8301	19.1145
	3–1	0.63	0.0721	0.8657	22.7968	2.20	0.1131	0.8287	19.5381
	3–2	0.80	0.0691	0.8742	23.2689	2.23	0.0937	0.8523	20.5691
	3–3	0.96	0.0717	0.8736	23.0275	2.47	0.0911	0.8548	20.6615
	3–4	1.12	0.0726	0.8701	22.9430	2.84	0.0946	0.8512	20.4794
	4–1	0.63	0.0695	0.8693	23.1950	2.25	0.1079	0.8325	19.7964
	4–2	0.82	0.0659	0.8769	23.6277	2.27	0.0865	0.8560	21.0037
	4–3	0.96	0.0668	0.8781	23.5622	2.52	0.0823	0.8615	21.2054
	4–4	1.12	0.0657	0.8777	23.7146	2.98	0.0814	0.8610	21.3135

Table 7: Inference time for a 3-to-3 setup for different caching configurations.

Setup	Time (s)
NAMVIS	0.96
w/o KV-cache	1.06
w/o ProPE-cache	1.28

E Ablation Details

Ablation studies. We conduct ablation studies to isolate the main architectural and optimization choices of NAMVIS, as reported in Table 5. Due to the cost of full-scale training, all ablation variants are trained on a randomly sampled 10K subset of our filtered Objaverse dataset for 200 epochs and evaluated on a held-out subset. These experiments are intended to compare design choices under a fixed compute budget.

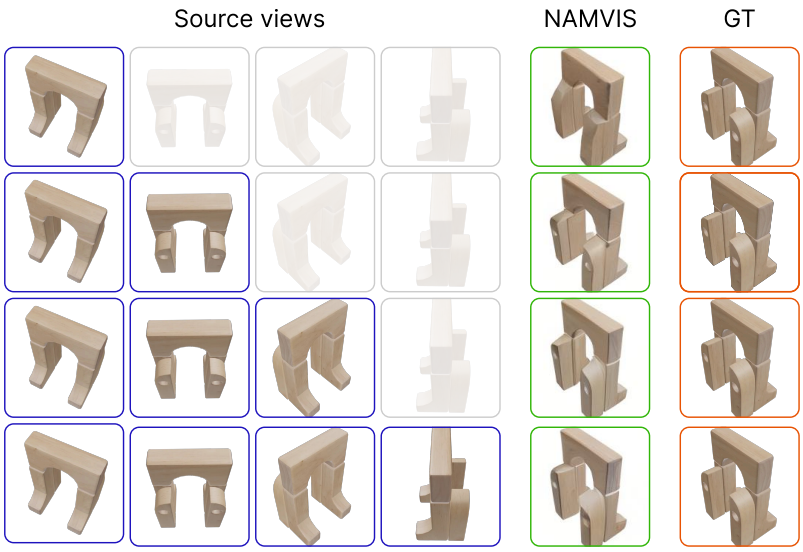

Figure 5: *Varying number of source views from 1 to 4. When views covering previously unseen areas appear, the generated image quality improves.*

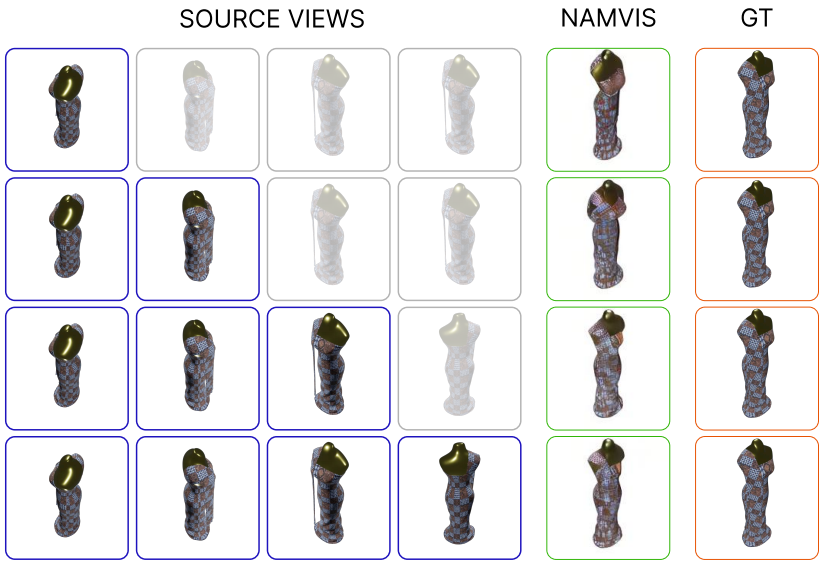

Figure 6: *Varying number of source views from 1 to 4. When views covering previously unseen areas appear, the generated image quality improves.*

E.1 Camera Pose Parameterization

Effective camera pose parameterization is critical for multi-view generation, as the generated images must strictly adhere to the target geometry. We compare our relative geometric approach against absolute pose conditioning via Plücker raymaps [Zhang et al., 2024]. In the raymap baseline, explicitly computed 3D rays are concatenated directly along the feature dimension. As shown in Table 5, injecting 3D camera geometry directly into the attention mechanism via multiscale ProPE [Li et al., 2025] outperforms the Plücker raymap conditioning across all metrics. This suggests that injecting geometry into the attention mechanism is more effective in our setting than concatenating explicit ray features.

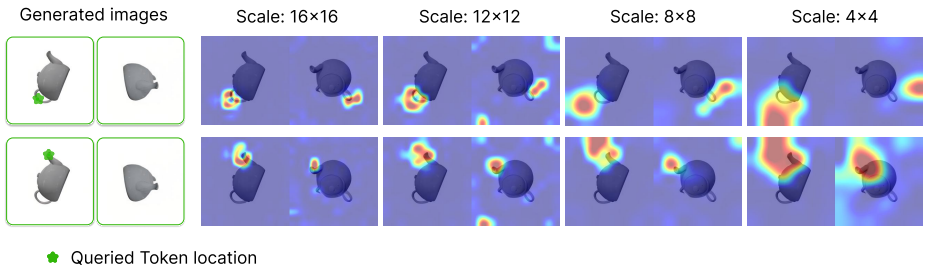

Figure 7: *Target-view self-attention pattern between two generated views.* For a queried token in the first target view, self-attention places high weight on corresponding regions in both generated target views, suggesting that target-target attention supports multi-view consistency.

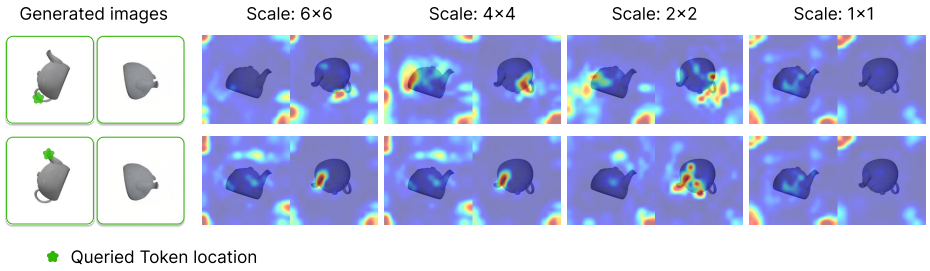

Figure 8: *Cross-attention pattern from a target image to two source images, layer 14.* Cross-attention assigns high weights to source-image regions that are visually aligned with the queried target location.

E.2 Reference Image Conditioning

Our geometry-aware reference conditioning operates via two distinct mechanisms: globally modulating the transformer blocks and initializing the sequence using a pooled `[SOS]` token, and aggregating fine-grained spatial details via dense cross-attention. We ablate both the source of these conditioning signals and the way they are injected into the autoregressive transformer.

Global Initialization (`[SOS]` Token). We ablate the source of the global semantic conditioning by replacing our VQ-VAE attentive pooling with pre-trained CLIP features [Radford et al., 2021]. While CLIP provides strong semantic priors, the VQ-VAE pooled features yield superior perceptual quality. Because the target sequence is composed of VQ-VAE tokens, deriving the `[SOS]` token from the same latent space provides better stylistic alignment and eliminates the domain gap between condition and target.

Dense Feature Aggregation. To understand the importance of high-frequency spatial conditioning, we completely remove the dense cross-attention, relying solely on the `[SOS]` token. This causes a substantial degradation in quality, suggesting that a single global token lacks the capacity to guide fine-grained image-to-image translation.

Next, we evaluate the injection mechanism itself. Replacing our dual-attention design (self-attention followed by cross-attention) with standard prefix conditioning, where source tokens are simply prepended to the target sequence, leads to a large performance drop. We hypothesize that directly prefixing dense source tokens interferes with the scale-structured autoregressive sequence, whereas cross-attention keeps source features in a separate key-value space and allows the target sequence to preserve its coarse-to-fine structure.

Finally, we ablate the source of the dense features by training a separate, dedicated feature extractor ("Learned Dense Features") via ConvNeXt [Liu et al., 2022] instead of utilizing the frozen VQ-VAE representations. The frozen VQ-VAE not only avoids the computational overhead of an auxiliary encoder but also performs marginally better, further validating the efficiency of our shared latent space design.

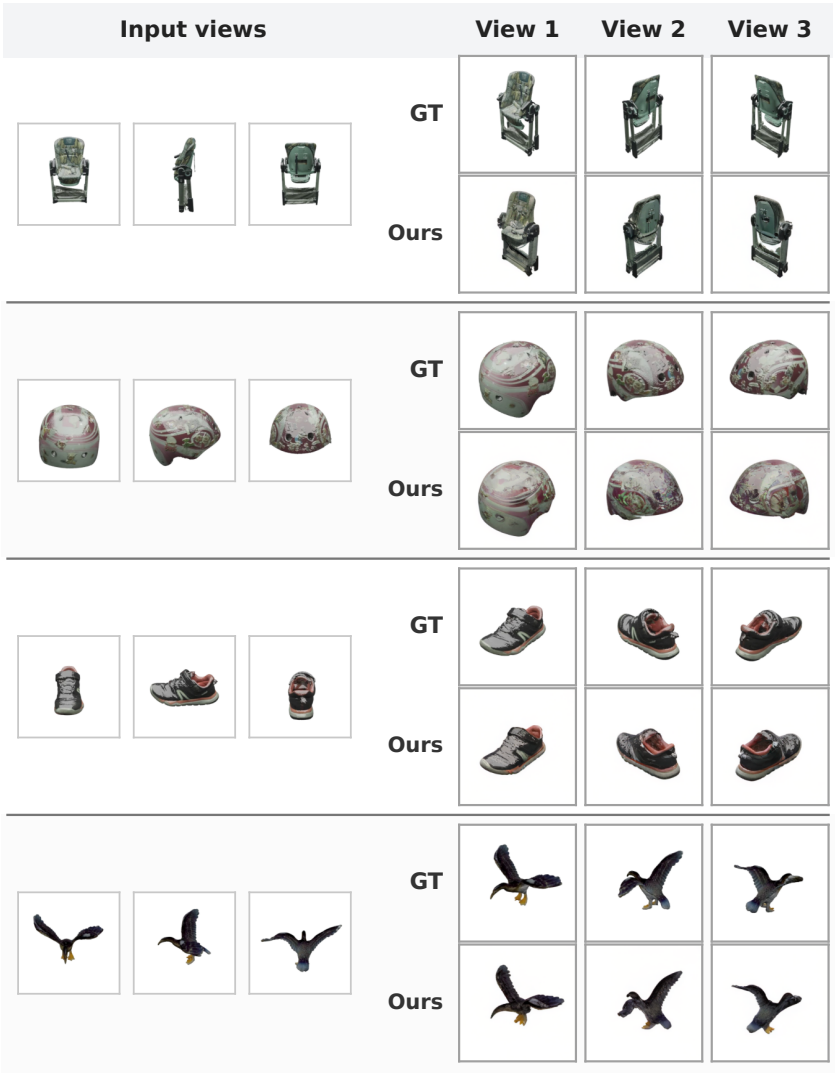

Figure 9: *Qualitative results of our method on OO3D [Wu et al., 2023].*

F Additional Qualitative Results

Figures 9 and 10 show additional NAMVIS generations across diverse object categories, shapes, poses, and textures on OO3D and GSO.

Scale progression. In Figure 11, we show how details appear on the final generated image as additional scales are summed up. Early scales 1×1 and 2×2 do not contribute significantly, the real final image contours start appearing after scale 4×4 is added. The image then progressively accumulates finer details at later scales.

G Attention Maps Analysis

We demonstrate comprehensive self- and cross-attention patterns across several layers and scales in Figures 12, 13, and 14. These visualizations suggest that NAMVIS learns correspondence patterns consistent with the underlying camera geometry.

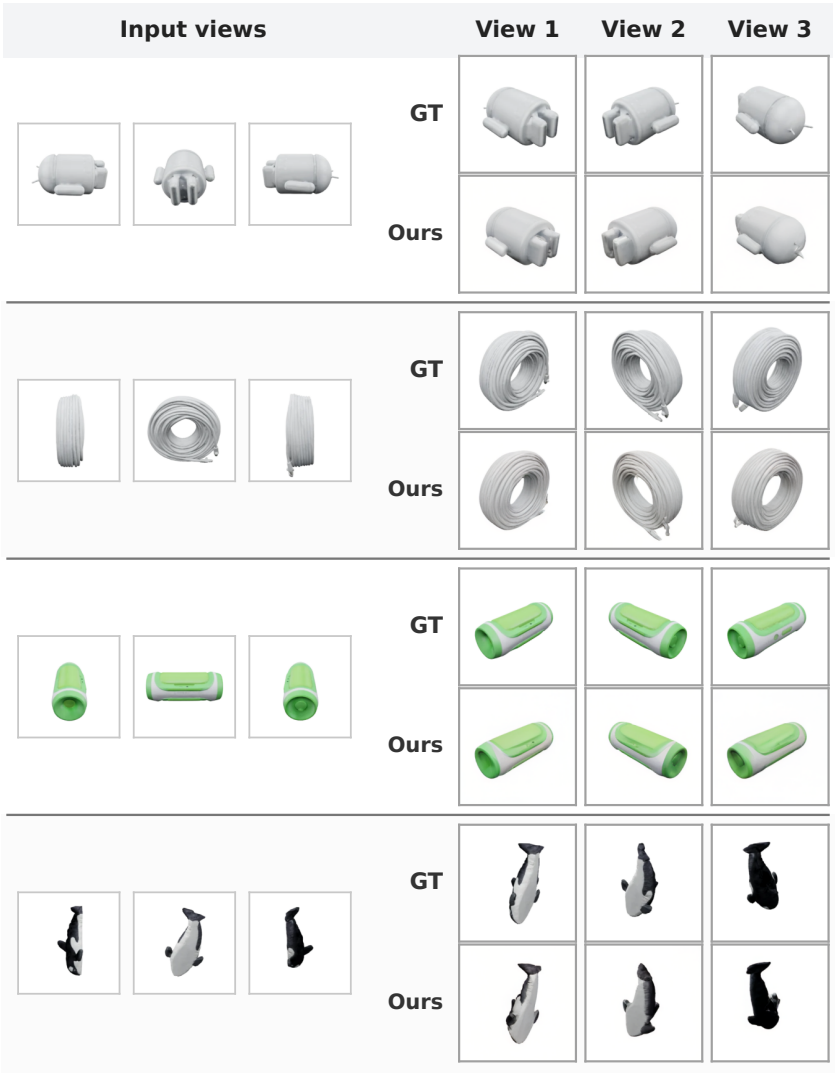

Figure 10: *Qualitative results of our method on GSO [Downs et al., 2022].*

H Evaluation Scenes

For reproducibility, we list the exact scenes used for evaluation on OmniObject3D (OO3D), GSO, and Objaverse. All methods are evaluated on the same scene list. Source and target camera configurations are fixed across methods whenever supported; for methods with camera-protocol constraints, we follow the protocol described in the main text.

I Broader Impacts

Our work on NAMVIS presents both positive and negative potential societal impacts. On the positive side, by providing a highly efficient, diffusion-free alternative for multi-view generation, our method significantly lowers the computational barrier and energy footprint required for 3D content creation. This has the potential to democratize applications in augmented and virtual reality, education, and digital design. Conversely, the ability to rapidly generate realistic, geometrically consistent multi-view content from sparse inputs carries potential risks. These include the facilitation of deceptive 3D media, the unauthorized replication of copyrighted intellectual property or artist styles, and the potential generation of harmful digital assets. Addressing these negative implications will necessitate ongoing community efforts to develop robust 3D watermarking, provenance tracking, and responsible deployment guidelines.

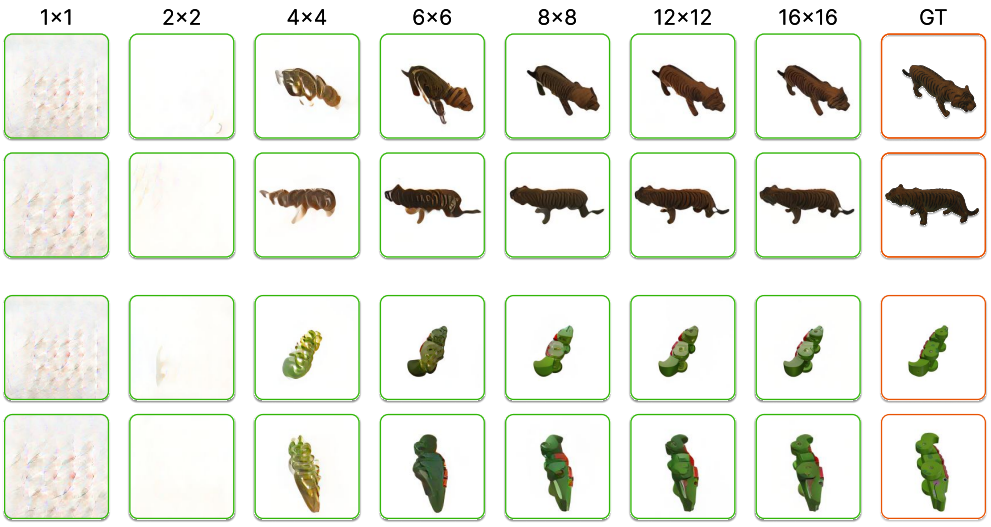

Figure 11: Visualization of image progression as new scales are summed up.

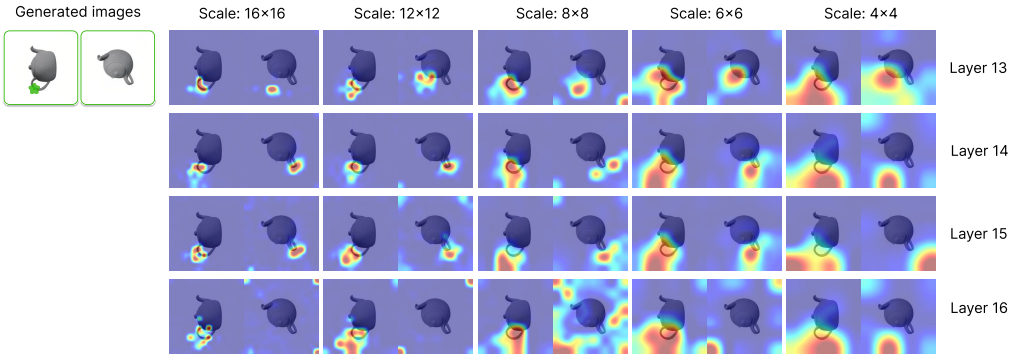

Figure 12: Visualization of self-attention pattern on layers 13-16, scales 4×4, 6×6, 8×8, 12×12, 16×16.

Table 8: Evaluation scenes from OmniObject3D (OO3D).

Scene	Scene	Scene
chair_002	chair_009	chair_017
chair_026	backpack_003	backpack_016
backpack_022	backpack_033	helmet_004
helmet_011	helmet_013	kettle_002
kettle_008	kettle_017	toy_animals_007
toy_animals_025	toy_animals_049	toy_animals_071
apple_006	apple_024	apple_058
vase_005	vase_012	vase_016
bottle_001	bottle_030	bottle_032
shoe_002	shoe_003	shoe_004

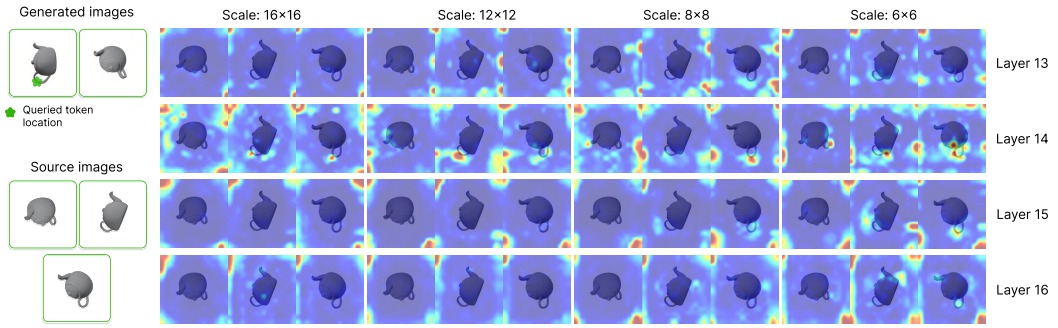

Figure 13: Visualization of cross-attention pattern on layers 13-16, scales 6×6, 8×8, 12×12, 16×16.

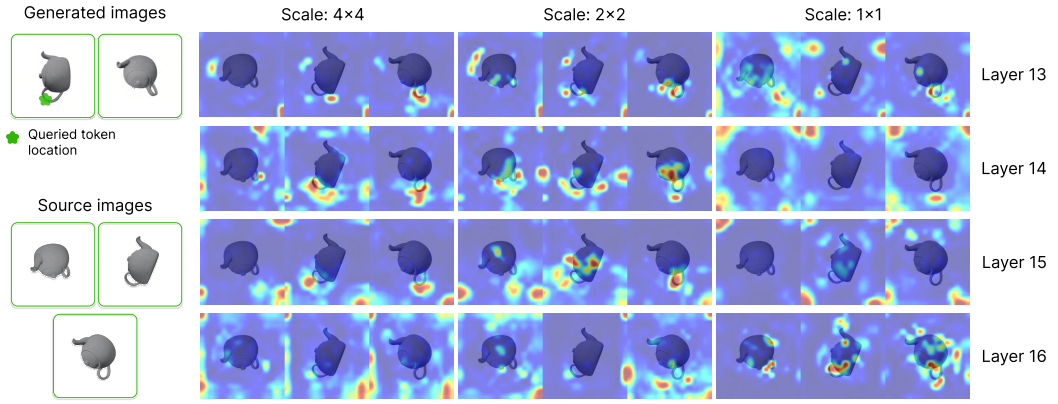

Figure 14: Visualization of cross-attention pattern on layers 13-16, scales 1×1, 2×2, 4×4.

Table 9: Evaluation scenes from GSO.

Scene	Scene
3D_Dollhouse_Swing	Netgear_Nighthawk_X6_AC3200_TriBand_Gigabit_Wireless_Router
Hasbro_Dont_Wake_Daddy_Board_Game	BUILD_A_ZOO
LEGO_5887_Dino_Defense_HQ	Melissa_Doug_Traffic_Signs_and_Vehicles
New_Super_Mario_BrosWii_Wii_Game	ALPHABET_AZ_GRADIENT
Crayola_Crayons_120_crayons	Tag_Dishtowel_Basket_Weave_Red_18_x_26
VANS_FIRE_ROASTED_VEGGIE_CRACKERS_GLUTEN_FREE	Office_Depot_HP_920XL_920_High_Yield_Black_and_Standard_CMY_Color_Ink_Cartridges
Android_Figure_Chrome	Black_Decker_Stainless_Steel_Toaster_4_Slice
Metallic_Gold_Tieks_Italian_Leather_Ballet_Flats	Canon_225226_Ink_Cartridges_BlackColor_Cyan_Magenta_Yellow_6_count
Seagate_1TB_Backup_Plus_portable_drive_Silver	Asus_Sabertooth_Z97_MARK_1_Motherboard_ATX_LGA1150_Socket
Playmates_nickelodeon_teenage_mutant_ninja_turtles_shredder	Schleich_Bald_Eagle
Philips_60ct_Warm_White_LED_Smooth_Mini_String_Lights	Shaxon_100_Molded_Category_6_RJ45RJ45_Shielded_Patch_Cord_White
SpiderMan_Titan_Hero_12Inch_Action_Figure_5Hnn4mtkFsP	JBL_Charge_Speaker_portable_wireless_wired_Green
Animal_Planet_Foam_2Headed_Dragon	Schleich_Spinosaurus_Action_Figure
Down_To_Earth_Orchid_Pot_Ceramic_Lime	Tory_Burch_Kiernan_Riding_Boot
Sootheze_Toasty_Orca	Mens_Authentic_Original_Boat_Shoe_in_Navy_Leather_RpT4GvUXRRP

Table 10: Evaluation scenes from Objaverse.

Scene	Scene	Scene
000074a334c541878360457c672b6c2e	001cfadfb9204424bccc45501ce6b90e	0023717f4f564cc99f4ded70db04f590
003199cc6ff2410cb2d8e6f8a9cbb163	003ebdf86df345d39dc166563229fb85	0051724d6efa42de84aaf8467629160f
0002c6eafa154e8bb08ebafb715a8d46	001d1b57e9df4273bede948b26429429	0023b3edbc114be188ca9d8f729dfaaf
003219b0bec442d29725847969e4b6bf	003ed2834130466da3a51bb7fdd6bde5	005f5630a54b442291aeb0a5d487353b
000b76f2b03e44e8ab44e1a1614be0f4	001fe8adb25a49d0b2650a5401dde019	0025c5e2333949feb1db259d4ff08dbe
0032696f5871429fbd0549d9628f812c	0046f208ef8d4988ba7bb9d297f29ec7	00616f328a8b4b5a8e689f61e70758b6
00124bcf3ca3463fbe05f28218cc0f5c	0022a3197f9646acbb9041eff2d1f55c	0025d57953fc4c8a80a44e59294e6841
0033322379a24798a6875a5cb2de54f5	0048e8224b174b759771e39ff521ee2e	0061788e0741400c82289337a24af4f6
00184eec45fe45ffa3826e9202fe7306	0022d2e01b014d328294b828e48defa1	00286954e2d54db8bc7832cc8682b6ff
00380c3f5cf548c9846faf3c42dfd6db	004d02243a5b4117afc4baa45eb1eba0	0064add4992b426cb2f862e5875ebf6d

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions of the paper are listed at the end of the introduction section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included limitations in the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include theoretical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All important information needed to reproduce the main experiments are provided in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We publicly release the curated Objaverse-XL dataset and source code, together with instructions for data preparation, training, and evaluation.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides all necessary training and test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report average performance across fixed evaluation configurations but do not provide confidence intervals, error bars, or formal statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides all necessary details about computer resources required to run our method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and believe that conducted research in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A section in Supplementary Materials is dedicated to broader impacts of our work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not work with data or models that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In our paper, we always properly cite the works or data mentioned in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have publicly released the curated Objaverse-XL dataset and source code, together with usage instructions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not include crowdsourcing experiments nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not include research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our paper does not use LLMs in our proposed method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.