Inference-Time Scaling for 3-D Diffusion Models

Tessa Everett[1] Tim Neumann[1]

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Abstract

Diffusion models now dominate 3D mesh synthesis from images because they can flexibly sample from complex, high-dimensional shape distributions. However, even the best performing 3D diffusion models still suffer from artifacts, poor fidelity, and perspective inconsistencies. Motivated by the substantial improvements observed in LLMs when allocating additional compute at inference-time, we apply inference-time-scaling algorithms tailored to 3D diffusion models to improve the quality of 3D generations. We propose a dynamic *Best-of-N* scaling algorithm for selecting candidate multiview generations, and employ a multimodal LLM grader as a general, customizable verifier. We conduct an experiment on a subset of the Objaverse mesh dataset and show that our scaling algorithm increases aesthetic quality and fidelity to ground truth meshes while offering interpretable justifications during verification.

1 Introduction

One key advantage of diffusion models is that increasing the number of denoising steps often improves sample quality [6]. But beyond a certain point, additional steps yield diminishing—and sometimes negative—returns [6], an observation we corroborate in the realm of 3D diffusion models (Figure 1). Existing algorithms for scaling diffusion models at inference-time have yet to be tested on 3D diffusion models, and require domain-adaption to perform optimally on this new class of model. This invites new methods for scaling 3D diffusion models. In this work, we propose an inference-time scaling approach for 3D diffusion models that combines (1) an adaptive Best-of-N sampling strategy and (2) a multimodal LLM grader that selects optimal multiview generations to improve downstream 3D reconstruction.

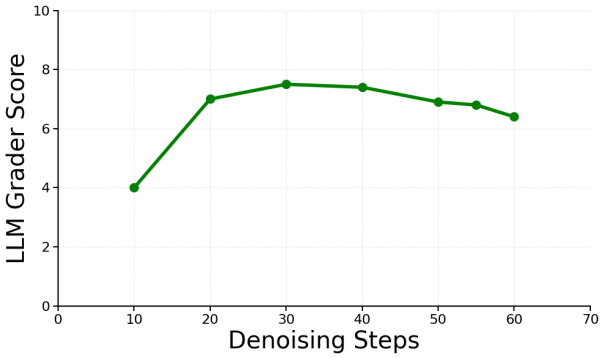

Figure 1: We generate 50 multi-view outputs from the diffusion pipeline of InstantMesh at various denoising steps, and extract LLM grader scores from the multimodal model Gemini. We observe plateauing and even degenerating multi-view quality, inviting better inference-time scaling algorithms

1.1 Inference-Time Scaling

Inference-time scaling strategies are algorithms that can be applied at inference-time to bolster model performance without updating model parameters, which requires costly retraining. Inspired by the rapid improvements to LLMs via inference-time scaling algorithms, we observe that analogous methods can be applied to 3D diffusion models, taking advantage of the inherent randomness present in transporting noisy samples to clean, novel generations. Unlike LLMs, diffusion models first sample gaussian noise, and repeatedly take steps to stochastically denoise the sample towards the target data distribution. As shown by Maq et al. [6]., image diffusion models follow impressive scaling laws when sampling N additional starting noises and selecting the best candidate according to a *verifier*, a scaling strategy coined *Best-of-N*. This algorithm has led to substantial improvements in aesthetic quality, prompt adherence, and overall coherence of 2D generations - motivating us to reformulate this approach to be compatible with 3D diffusion models. We propose a variable N that adaptively adjusts the number of starting seeds searched over at inference-time, reducing time and compute for easy generation tasks and

reallocating them towards difficult generations. We furthermore employ a multimodal LLM grader that has the capacity for assessing multi-view candidates on arbitrary prompting and reweighting. Our contribution consists of

1. *Adaptive Best-of-N* sampling strategy that resamples starting initial Gaussian noises from a generative model *up to N times* based on feedback from the LLM verifier, reallocating compute from "easy" to "difficult" reconstructions.

2. *Multi-modal LLM grader* prompted to evaluate the output of the diffusion pipeline with a score of 1-10. The LLM can receive arbitrary prompting for evaluating candidate multi-view generations, and outputs interpretable natural language justifications alongside scores.

1.2 Motivation

Diffusion models are a class of generative models trained to learn a backwards denoising process. At generation time, they start from pure random noise and iteratively apply these learned denoising steps to sculpt novel samples. They have quickly dominated image, video, audio, and more recently 3D reconstruction domains, due to their capacity for sampling from high dimensional complex data distributions, which are common in these visual domains. 3D diffusion models in particular have revolutionized fields like interior design and manufacturing, where obtaining high fidelity and aesthetic 3D models from a single input image can greatly reduce labor costs and expedite commercial processes. However, perspective inconsistencies, artifacts, and poor fidelity still plague these generated models and necessitate costly human verification, inviting research into scaling the robustness and quality of 3D diffusion models.

1.3 Related Works

1.3.1 Diffusion Models for 3D Generation

Diffusion models rapidly accelerated the field of 3D object generation by allowing reconstruction models to condition on 2D, multiview priors. DreamFusion [7] was one of the first models to utilize a two-step pipeline for 3D object synthesis, first passing a single image into a multi-view diffusion model that outputs novel perspectives of the object in the original image. Then, a neural radiance reconstructs the 3D object using the multiview generations as guidance. This diffusion-reconstruction pipeline has largely become the state of the art for 3D object generation, with models like Zero-123 achieving improvements by further conditioning on changes in multiview images. Despite these advancements, there is still a large potential for additional improvements by scaling at inference time.

1.3.2 Inference Time Scaling for Diffusion Models

A natural lever for improving diffusion outputs is to increase the number of denoising steps, yet several studies report diminishing or even negative returns beyond a dataset-specific horizon. Complementary to altering denoising steps, a line of work explores *search-based* scaling strategies at inference time. Best-of-N sampling (randomly drawing N noise seeds and selecting the highest-reward sample) consistently improves FID and aesthetic metrics for text-to-image diffusion transformers.

1.3.3 LLM Graders

Human evaluation of 3-D outputs is costly; recent work therefore leverages LLMs as reference-free judges. Vision-language models like GPT-4V have achieved medical image assessments on par with human experts [4], demonstrating their ability to interpret high-resolution spatial features , highlighting their capacity to reason over high-resolution spatial detail. However, prior graders operate on 2-D projections; our multimodal rubric explicitly incorporates multi-view coherence and mesh fidelity, bridging the gap between 2-D image assessment and full 3-D reconstruction quality.

1.4 Summary

Taken together, existing literature shows that (i) sampling diversity via noise-seed search, (ii) LLM-based evaluators each improve generative fidelity in isolation, and we furthermore observe that adaptive compute allocation can be helpful in 3D domains when the difficulty of reconstruction tasks vary greatly. Our contribution unifies these strands by integrating an adaptive *Best-of-N* search with a multimodal LLM grader, yielding significant improvements on samples from the Objaverse dataset without retraining the diffusion backbone.

2 Method

We will focus our analysis on 3D diffusion models that employ a two stage, diffusion-reconstruction pipeline,

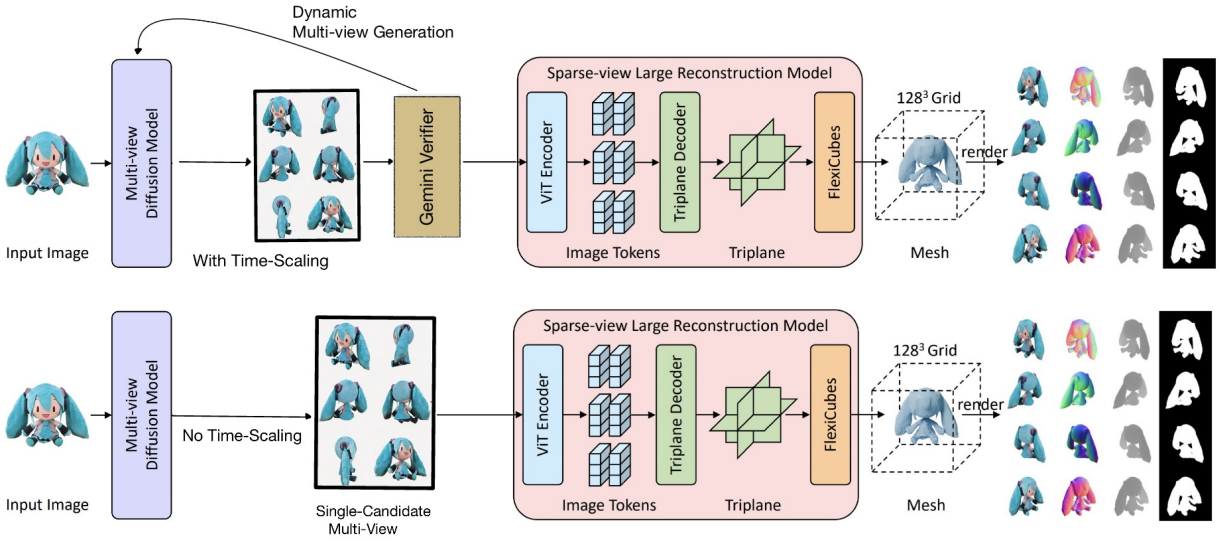

Figure 2: Top: Our modified InstantMesh pipeline, which uses a multimodal LLM grader (Google Gemini) to screen multiview outputs against a score threshold before 3D reconstruction. Bottom: The standard InstantMesh model without inference-time scaling.

as modeled in Figure 2. Our inference-time scaling algorithm will aim to scale the quality of the multi-view outputs from the diffusion stage with the aim of improving downstream 3D generation fidelity and aesthetic quality. Our framework can be understood as having two scaling axes: the search algorithm (*Adapative Best-of-N* and verifier (LLM Grader). We discuss these components below.

2.1 Adaptive Best-of-N

Under the observation that some starting noises lead to higher quality generations, *Best-of-N* has been shown to be a simple, effective scaling algorithm for image diffusion models. We furthermore observe that there is a more concrete notion of the ground truth in 3D reconstruction that is partially revealed by the input image, which in turn greatly varies the difficulty of the reconstruction task. We therefore adaptively change the number of regenerations based on a score threshold. N is now an *upper bound* on the number of generations, with the score threshold being in place to shortcircuit the algorithm if feedback from the LLM suggest that the multiview generations already compose a promising input to the 3D reconstruction model. Our algorithm is concretely outlined in Appendix B.

2.2 LLM Grader

The *adaptive Best-of-N* algorithm's efficacy largely relies on the validity of the verification step. Rather than relying on pretrained task-specific networks to act as verifiers—which might generalize poorly between diffusion models, input images, and ground truth objects while exhibiting limited customization—we rely on a more general joint text-image model in the form of a multimodal LLM. Using an LLM grader as a verifier allows for customizable prompting and scoring, a wide range of content in the input images, and interpretable explanations for scores.

Formally, we give the LLM grader a JSON of the input image, the group of multi-view generations, and a prompt outlining what metrics to assess the multi-view generation on, requiring grading on a scale of 1 to 10. In our experiment, we prompted the LLM to assess on the basis of *(1) Consistency with Input Image, (2) Aesthetic Quality, (3) Visual Consistency across Views*. Additionally, each of these sub scores can be weighted with *low, medium, high* importance. The final outputscore is a weighted average of the subscores as a proxy for 3D reconstruction potential. Our exact prompt can be found in Appendix A.

2.3 Experiment

We conduct the following experiment using the 3D diffusion model InstantMesh for its fast, low memory

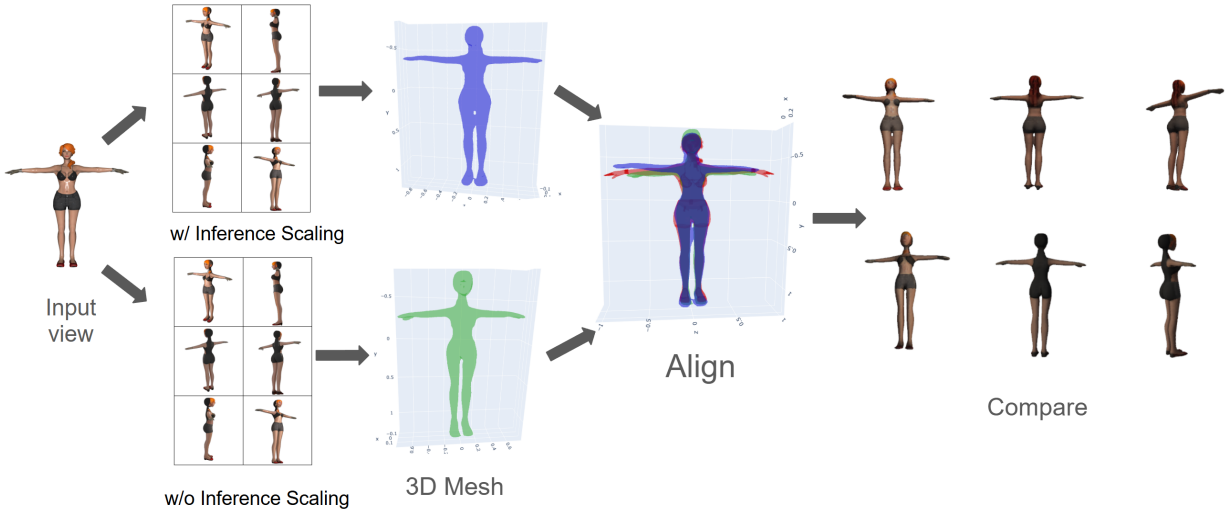

Figure 3: Outline of our expreimental process. For each of the 145 sampled meshed, we input a 2D view to the base and inference time-scaled model which genereated meshes. We then align the meshes using ICP and take 32 views of each generated mesh to compare consistent views of generated meshes against ground truth viewss.

overhead 2 step diffusion pipeline. We first extract input views from ground truth meshes taken from a subset of the Objaverse mesh database to create a dataset of 145 (input image, ground truth mesh) pairs. We generate meshes both with and without adaptive Best-of-N (with a threshold value of 8), and assess their fidelity to the ground truth. For 3D geometric evaluation, we use the Chamfer Distance (CD) to measure the 3D structural similarity between the generated mesh and ground truth [3]. For 2D perceptual evaluation, we render 32 novel views of both the generated and ground truth mesh and report the LPIPS, PSNR, SSIN, and FS scores. Importantly, we write our own 3D alignment and 2D view extraction algorithms and evaluate on a different dataset, so our reported metrics are not comparable with what was reported in Xu et al. [8].

3 Results

Quantitative Results. To compare generated meshes against the ground truth, we first normalize all meshes to fit within a $[-1, 1]^3$ cube and align them using an Iterative Closest Point (ICP) algorithm [2] (see Appendix C for implementation details). For each object in our dataset, we generate 32 consistently sampled camera views, evenly distributed on a spherical surface around the aligned mesh. We take 2D images of the same views of the ground truth, the base model output, and our inference-time-scaled model output. We then compare generated views with ground truth for each 2D metric (PSNR, SSIM, LPIPS, FS) by first averaging across the 32 values of each metric for each generated mesh, and then reporting the average over all 145 objects in the dataset. We achieve significant improvement across all 2D metrics, with values shown in Table 1. We observe a 1.647 increase in PSNR which measures the pixel-wise fidelity between two images. A higher PSNR indicates our novel views are closer to the ground truth image in terms of pixel-wise accuracy when compared with the base model. Higher PSNR does not always mean better perceptual quality to the human eye. For instance, PSNR is blind to perceptual details like edges or local contrast and two images can have high PSNR but differ in texture or structure. LPIPS, which our model saw a 0.0855 decrease in (where lower is better), addresses these, and quantifies how visually similar two images are based on the activations of a pre-trained deep neural network (in our case AlexNet [5]). Substantially outperforming the base model on these two 2D metrics suggests our infrerence time-scaled model is producing images that are both more numerically accurate and perceptually faithful to the ground truth than the base model, and is a strong indication that our model improves fidelity and perceptual quality. This claim is further reinforced by the marked improvement in SSIM and FS, which respectively capture structural similarity and distributional realism, offering additional evidence that our generated views are

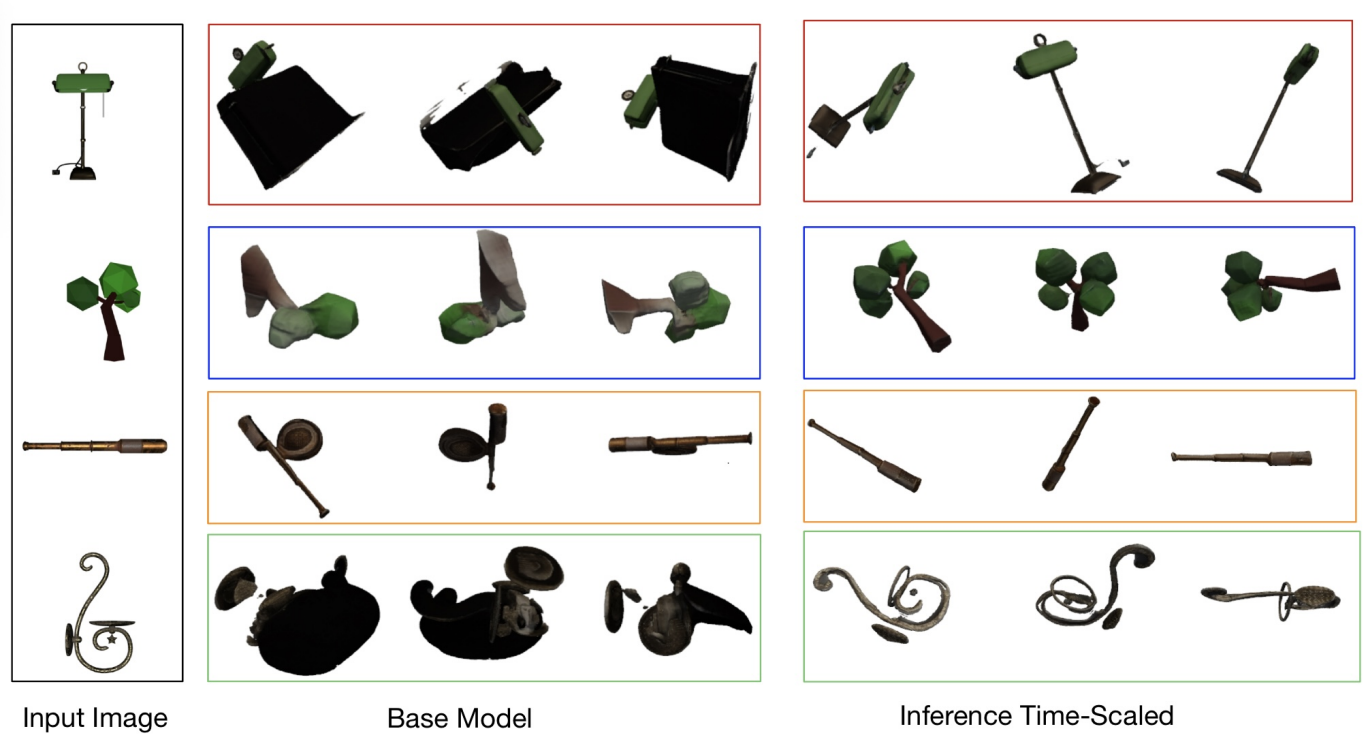

Figure 4: **Left:** Input images. **Middle:** Base InstantMesh reconstructions. **Right:** Inference-time-scaled reconstructions. Our inference-time-scaled model consistently reduces artifacts and improves both geometric fidelity and texture quality across diverse objects. In contrast to the blocky distortions and collapsed details seen in the base InstantMesh outputs, the scaled model produces cleaner multiviews that better preserve thin structures, fine curves, and surface shading (see lamp and wall sconce examples in the figure).

not only pixel-aligned but also semantically consistent and visually coherent.

To assess significance, we assumed independence between our 145 randomly sampled meshes from the Objaverse-XL dataset due to Objaverse-XL large size and object diversity (Objaverse-XL contains over one million objects) [1]. We first assessed normality of the paired differences using the Shapiro–Wilk test, and then applied paired t-tests (or a Wilcoxon signed-rank test when appropriate) for our 2D metrics (see Appendix D). In every case, $p \ll 0.001$, demonstrating that our inference-time scaling yields consistent and highly significant improvements across all four metrics.

We also evaluated geometric fidelity via the Chamfer Distance (CD) between predicted and ground-truth point clouds. The paired differences failed the Shapiro–Wilk normality test ($W = 0.931$, $p < 0.001$), so we report both parametric and nonparametric results. The mean CD improvement was 0.0262 (Gemini − baseline), but this did not reach significance: the paired t-test yielded $t = -0.508$, $p = 0.613$, and the Wilcoxon signed-rank test gave $W = 8068$, $p = 0.812$. Thus, although the average Chamfer Distance was slightly lower with our model, the difference is not statistically significant.

Table 1: Quantitative Results on 145 Objaverse-XL Objects

Method	PSNR↑	SSIM↑	LPIPS↓	CD↓	FS↑
IM (base)	10.882	0.7148	0.4094	0.2342	0.8587
IM (ours)	**12.529**	**0.7431**	**0.3239**	**0.2080**	**0.8774**

Qualitative Results. Figure 4 compares reconstructions from the base InstantMesh model (middle column) and our inference-time-scaled model (right column), alongside the original input images (left column). Across all four examples, our scaled model consistently reduces obvious artifacts and more faithfully reproduces both geometry and texture. Overall, these examples illustrate that adaptive Best-of-N sampling with an LLM grader not only improves quantitative metrics but also yields visually cleaner, more artifact-free multiview sets—critical for high-quality

3D reconstruction.

4 Conclusion

In this work, we introduced a model-agnostic, customizable inference-time scaling algorithm for 3D diffusion models that combines adaptive Best-of-N sampling with a multimodal LLM grader. By dynamically allocating compute based on generation difficulty and leveraging a reference-free evaluator to select high-quality multiview outputs, our method significantly improved 2D perceptual metrics (PSNR, SSIM, LPIPS, FS) and yielded a modest average reduction in Chamfer Distance, though the latter did not reach statistical significance on our 145-sample set. Notably, in several cases the gains achieved through scaling rivaled those reported from fundamental architectural advancements in 3D diffusion models, highlighting inference-time scaling as an untapped avenue for enhancing generation quality. Crucially, these improvements were obtained without any retraining of the diffusion backbone, underscoring the practicality and compute-efficiency of our approach.

5 Discussion

While our results demonstrate clear benefits of inference-time scaling in visual fidelity and perceptual quality, several limitations merit consideration:

- **Dataset size:** Our evaluation was limited to 145 Objaverse-XL meshes due to compute constraints (7hours of A100 time), which may underpower detection of small geometric gains such as Chamfer Distance improvements.
- **Chamfer Distance significance:** Although the mean CD decreased by 0.0262, paired tests did not reach significance, suggesting that larger or more diverse samples are needed to validate geometric improvements.
- **Black-box verifier:** The multimodal LLM grader operates as an external API (e.g., Gemini) with opaque internal weighting, which may limit reproducibility and introduce domain biases.

Future Work. To place our findings in context of prior 3D diffusion research, we will:

- Conduct evaluations over a larger, unified InstantMesh dataset, enabling direct comparison with existing architectural benchmarks.
- Explore distillation of LLM feedback into lightweight internal verifiers to reduce external API dependence.
- Investigate more fine-grained adaptive schedules for compute allocation across varied object categories and reconstruction difficulties.

Overall, our project reveals that inference-time scaling algorithms—long exploited in LLMs—represent a promising, underexplored direction for advancing 3D generative modeling without incurring the cost of backbone retraining.

References

[1] Allen Institute for AI. Objaverse-XL: A Universe of 10M + 3D Objects. `https://objaverse.allenai.org/`

[2] Besl, P. J. and N. D. McKay. A method for registration of 3–D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[3] Fan, H., Su, H., and Guibas, L. J. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 246–254.

[4] Li, Y., Liu, Y., Wang, Z., Liang, X., Wang, L., Liu, L., Cui, L., Tu, Z., Wang, L., and Zhou, L. A Systematic Evaluation of GPT-4V's Multimodal Capability for Medical Image Analysis. *arXiv preprint arXiv:2310.20381*, 2023.

[5] Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[6] Maq, N., Tongq, S., Jia, H., Hu, H., Su, Y.-C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., and Xie, S. Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps. *arXiv preprint arXiv:2501.09732*, 2025.

[7] Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[8] Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., and Shan, Y. InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models. *arXiv preprint arXiv:2404.07191*, 2024.

A LLM Grader Prompt

```
You are an expert 3D artist and image quality evaluator.
You will be provided with data in the following sequence:
1. This instructional prompt.
2. A section starting with "--- Initial Input Image ---" followed by a single base image.
This is the original image used to generate 3D views.
3. A section starting with "--- Candidate Multiview Set ---" followed by a set of 6
multiview images. These are candidate views for reconstructing a 3D model.

Your task is to evaluate the "Candidate Multiview Set" based on its quality and suitability for
3D reconstruction, WITH CRITICAL CONSIDERATION of its consistency with the "Initial Input Image".

Evaluation Criteria for the "Candidate Multiview Set":
1.  **Consistency with Initial Input Image (Weight: High)**:
    *   Object Identity: Do the multiviews clearly depict the same object/subject as shown in
    the initial input image?
    *   Key Features: Are important details, shapes, and characteristics from the input image
    accurately represented in the multiviews?
    *   Style and Appearance: Is the artistic style, texture, and overall appearance consistent
    between the input image and the multiview set?
2.  **Aesthetic Quality (Weight: Medium)**:
    *   Visual Appeal: Are the multiview images individually and collectively visually appealing?
    *   Clarity & Detail: Are the images sharp, well-defined, and free of excessive noise or
    artifacts?
3.  **Visual Consistency Across Views (Weight: Medium)**:
    *   Coherence: Do the 6 views look like they belong to the same 3D object viewed from
    different
    angles?
    *   Lighting & Shading: Is lighting and shading consistent across all views?
    *   Texture/Color Continuity: Do textures and colors flow smoothly and realistically from
    one view
    to another?
4.  **3D Reconstruction Potential (Weight: High)**:
    *   Completeness: Do the views provide sufficient information to reconstruct a coherent
     3D shape?
    *   View Diversity: Are the views diverse enough to capture different aspects of the object?
    *   Ambiguity: Are there minimal ambiguities or contradictions that would hinder 3D
    reconstruction?

Output Format:
Provide your evaluation as a single, valid JSON object. The JSON object should have the
following structure:

{
  "consistency_with_input_image": {
    "score": <number from 0 to 10, where 10 is perfect consistency>,
    "explanation": "<brief explanation for the score, max 2 sentences>"
  },
  "aesthetic_quality": {
    "score": <number from 0 to 10, where 10 is excellent>,
    "explanation": "<brief explanation, max 2 sentences>"
  },
  "visual_consistency_across_views": {
```

```
    "score": <number from 0 to 10, where 10 is perfect consistency>,
    "explanation": "<brief explanation, max 2 sentences>"
  },
  "reconstruction_potential": {
    "score": <number from 0 to 10, where 10 is excellent potential>,
    "explanation": "<brief explanation, max 2 sentences>"
  },
  "overall_score": <number from 0 to 10, this should be a weighted average or holistic
  assessment
  based on the above criteria, with higher weight on 'Consistency with Initial Input Image'
  and '3D
  Reconstruction Potential'>,
  "overall_assessment": "<a concise summary (2-3 sentences) of why this candidate set
  is good or
  bad, highlighting its consistency with the input image and overall 3D suitability>"
}

Important:
- Scores must be numbers (e.g., 7), not strings (e.g., "7/10").
- Ensure the entire response is ONLY the JSON object. Do not include any text before or
after the JSON.
```

B Adapative Best-of-N

Algorithm 1 Adaptive Best-of-N Sampling

Require: input image I, model M,N, score threshold τ

$n \leftarrow 0$, $best\_score \leftarrow -\infty$
while $n < N$ **and** $best\_score < \tau$ **do**
  sample noise seed s_n
  generate multi-view set $V_n \leftarrow M(I, s_n)$
  $(score_n, reason_n) \leftarrow \text{LLM\_GRADE}(I, V_n)$
  if $score_n > best\_score$ **then**
    $best\_views \leftarrow V_n$; $best\_score \leftarrow score_n$
  end if
  $n \leftarrow n + 1$
end while
return $best\_views$

C Implementation Details

- **Diffusion backbone:** denoising steps = 30, guidance scale = 7.5.
- **LLM grader:** Multimodal Gemini API (v2025-04).
- **Camera sampling:** 32 views via Fibonacci sphere.
- **ICP alignment parameters:**

Parameter	Value
n_iter	100
count_source	10,000
count_target	20,000
initial_transforms	None
fixed_scale	true
outliers	0.0
on_surface	false
min_scale	0.5
max_scale	2.0
coarse_iter	150
fine_iter	100
outliers	0.1
plot	false
test_reflections	true

Table 2: Open3D ICP alignment settings.

D Statistical Tests for 2D Metrics

Metric	Shapiro–Wilk		Significance Test	
	W	p	Test	Statistic (p)
PSNR	0.984	0.1041	Paired t	$t = 5.715$ (6.35e8)
SSIM	0.972	0.0060	Paired t	$t = 4.289$ (3.33e5)
			Wilcoxon	$W = 2840$ (8.32e6)
LPIPS	0.985	0.1346	Paired t	$t = -8.062$ (3.00e13)
FS	0.985	0.1343	Paired t	$t = 7.483$ (7.31e12)

Table 3: Normality (Shapiro–Wilk) and paired significance tests for our four 2D metrics.

Table 4: Per-instance Mean and Standard Deviation Across 145 Objaverse-XL Objects

Method	PSNR↑	SSIM↑	LPIPS↓	FS↑
IM (base)	10.88 ± 2.95	0.7148 ± 0.0915	0.4094 ± 0.1131	0.8587 ± 0.0452
IM (ours)	$\mathbf{12.53 \pm 3.43}$	$\mathbf{0.7431 \pm 0.0946}$	$\mathbf{0.3239 \pm 0.1224}$	$\mathbf{0.8774 \pm 0.0470}$