UMI3D: ROBUST 3D GENERATION ON UNCONSTRAINED MULTI-IMAGE INPUTS VIA SIMULTANEOUS FOCUSING CROSS-ATTENTION ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent 3D foundation models enable 3D asset generation from a single image, but degrade substantially when extended to unconstrained multi-image inputs, which comprise collections of images depicting the same object yet differing in pose, style, illumination, and local details. This degradation often results in distorted geometry, over-smoothed textures, and inconsistent colors. Our analysis suggests that a key limitation is the mismatch between single-image cross-attention and multi-image conditioning: existing models lack an explicit mechanism for selecting a reference image for each voxel at each denoising step. Based on this observation, we propose UMI3D, a training-free framework that restructures cross-attention by explicitly routing each voxel to its most informative image, which prevents conflicting visual evidence from being indiscriminately mixed during generation, thereby unlocking strong performance on inconsistent inputs. To realize this routing, UMI3D introduces *Simultaneous Focusing Cross-Attention* (SFC-Attn), which activates all conditioning images at each denoising step while allowing each voxel to focus on the single image that best explains it. We further derive the Voxel Reference Score (VRS), a model-intrinsic metric for voxel–image affinity that requires no external models. Extensive experiments show that UMI3D unlocks the multi-image potential of single-image 3D generation frameworks across diverse tasks. Code will be released.

1 INTRODUCTION

Generating 3D objects from multiple input images is a common approach for many real-world 3D workflows, such as creating 3D characters from multi-view concept art and constructing virtual environments from collections of scene photographs. Multiple input images provide complementary observations that reduce single-view ambiguity and reveal otherwise occluded regions. However, traditional multi-view to 3D methods require known camera poses, consistent object details and illumination across views. Since image collections satisfying all these constraints are difficult to obtain in practice, we introduce *unconstrained multi-image 3D generation*, a new task that aims to generate a coherent 3D asset from multiple images without guaranteed camera poses or cross-view consistency. We refer to such image collections as *unconstrained multi-image inputs*.

A promising approach to this new task is to build upon recent 3D generative foundation models (Li et al., 2025c; Wu et al., 2025; Team, 2025; Xiang et al., 2025; Li et al., 2025b; Chen et al., 2025), which can generate high-quality 3D assets from a single image. However, these models are designed around a single conditioning image and struggle to integrate multiple unconstrained inputs. When naively extended to multi-image, they often produce distorted geometry, over-smoothed or inconsistent textures across views. This limitation motivates us to investigate how their strong 3D priors can be retained while effectively incorporating information from multiple unconstrained images.

Some existing methods (Team, 2025; Li et al., 2025b) incorporate multi-image conditioning during training, but are designed for relatively consistent multi-view observations. When the inputs exhibit substantial discrepancies in visibility, appearance, or local details, these methods produce entangled 3D structures that limit their real-world applicability. Meanwhile, single-image 3D generation frameworks (Xiang et al., 2024; 2025; Wu et al., 2025) can be naively extended to multiple inputs,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

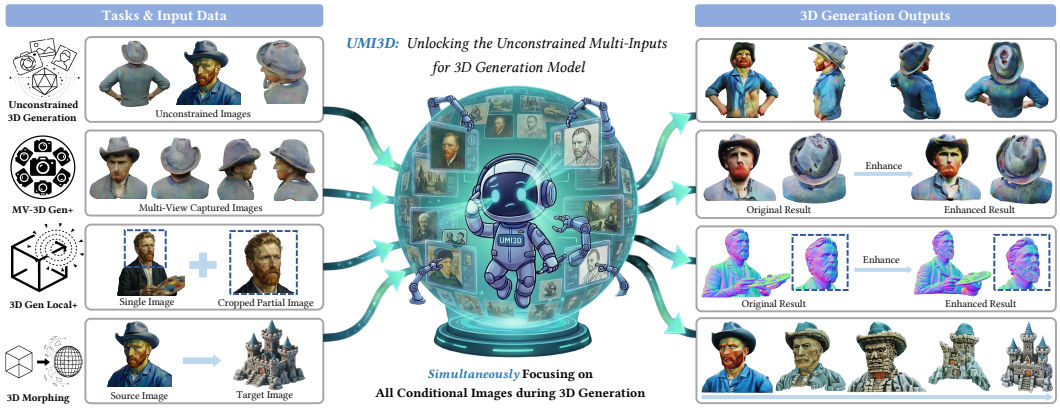

Figure 1: UMI3D can generate a coherent 3D asset with unconstrained multi-image input and support multiple downstream applications in a training-free manner.

for example by alternating the conditioning image across denoising steps Xiang et al. (2024) or by combining independently generated outputs Xiang et al. (2025). However, such strategies force the model to switch its conditioning reference across denoising steps, often leading to detail loss, conditioning oscillation and noticeable discrepancies across views.

Our analysis (Sec. 4.3) shows that this gap traces back to the mismatch between single-image cross-attention design and the multi-image setting. Existing single-image 3D generation frameworks are built around cross-attention to a single conditioning image, but multi-image generation requires different parts of the 3D asset to extract information from different references. Without an explicit mechanism to regulate this process, visual cues from multiple images can interfere with one another, leading to unstable geometry and inconsistent textures.

In this work, we propose **UMI3D**, a training-free framework for unconstrained multi-image 3D generation. Its central idea is to recast cross-attention as a dynamic voxel-wise image routing problem. We show that explicitly routing each 3D voxel to its most informative image prevents conflicting visual cues from being indiscriminately mixed, while retaining access to all input images throughout generation. This strategy allows the strong 2D-3D priors of existing single-image foundation models to be effectively extended to unconstrained multi-image settings.

At its core, UMI3D introduces a **S**imultaneous **F**ocusing **C**ross-**A**ttention (SFC-Attn) module, which activates all conditioning images simultaneously at each denoising step while letting each 3D voxel focus on the image that best explains it, thereby avoiding the conditioning oscillation caused by switching references across steps. To drive this routing, we analyze the original cross-attention layers and derive a metric called **V**oxel **R**eference **S**core (VRS), which automatically establishes voxel–image correspondences without relying on any external models (e.g., 3D matching or segmentation networks). VRS exploits how the baseline 3D generation model utilizes both global and local image information, allowing reliable voxel-to-condition assignments in a purely model-intrinsic manner.

Beyond unconstrained multi-image 3D generation, the same mechanism extends naturally to several additional settings, including 3D morphing, multi/single-view 3D generation enhancement, as shown in Fig. 1, which we showcase as applications of our framework. To systematically evaluate the core task, we construct a benchmark of unconstrained multi-image inputs covering diverse cross-view inconsistencies. Extensive experiments demonstrate that UMI3D delivers strong and robust multi-image 3D generation performance. Our main contributions can be summarized as:

- We propose UMI3D, a training-free framework that adapts single-image 3D foundation models to unconstrained multi-image inputs without updating the backbone parameters.
- We propose Simultaneous Focusing Cross-Attention (SFC-Attn), which uses a model-intrinsic Voxel Reference Score (VRS) derived from original cross-attention to route each voxel to its most informative image, enabling consistent multi-image 3D generation.
- We have constructed a new benchmark for unconstrained multi-image 3D generation with realistic cross-view inconsistencies. Extensive experiments validate the effectiveness and robustness of UMI3D.

2 Related Works

3D Generative Foundation Models Recently, a growing body of works (Luo & Hu, 2021; Nichol et al., 2022; Hui et al., 2022; Müller et al., 2023; Chen et al., 2023; Shue et al., 2023; Poole et al., 2022; Wang et al., 2024) has explored the transfer of 2D diffusion models (Ho et al., 2020; Rombach et al., 2022) to various 3D generation tasks. To further improve generation quality and computational efficiency, another line of research (Ren et al., 2024; Zhang et al., 2023; 2024; Chen & Zhang, 2019; Hertz et al., 2022) investigates how to perform 3D generation in compact latent spaces. Most recently, the emergence of large-scale 3D generative models (Zhao et al., 2025; Xiang et al., 2024; Wu et al., 2024; Hui et al., 2024; Li et al., 2025b; Seed, 2025; Chen et al., 2025) has further accelerated progress in 3D generation, providing stronger and more versatile priors for learning 3D content. Voxel-based 3D generative models, exemplified by TRELLIS (Xiang et al., 2025) and Direct3D (Wu et al., 2025), have recently emerged as a main paradigm for 3D generative foundation models. However, due to their particular training schemes and their design focus on broadening practical application scenarios, these models are typically tailored for single-image conditioning.

Multi-View 3D Generation Pipeline Several approaches have explored 3D generation with multi-view inputs. Models such as TripoSG (Li et al., 2025b) and Hunyuan3D (Team, 2025) are optimized with multiple views of the object during the training stage. Some works (Feng et al., 2025; Cheng et al., 2025; Wen et al., 2025; Gao et al., 2025) utilize the pre-trained multi-view diffusion models for more input images, aiding the single-image generation task. However, the above multi-image input pipelines are restricted to consistent views of the same object for the fixed model parameters. Some voxel-based generation methods like TRELLIS (Xiang et al., 2024), TRELLIS.2 (Xiang et al., 2025) and Direct3D-S2 (Wu et al., 2025) adopt a simple strategy that switches the conditioning image at each denoising step to fuse information from multiple views. However, this often causes error accumulation and degrades the final 3D quality.

3D Generation from Unconstrained Inputs In 3D reconstruction, many works (Xu et al., 2024; Martin-Brualla et al., 2021; Kulhanek et al., 2024; Sabour et al., 2025; Lin et al., 2025; Go et al., 2025; Qu et al., 2024) have investigated scene reconstruction from wild, unconstrained sets of input images. Yet, in the context of 3D generation, such methods can only recover the portions of an object that are directly observed and are unable to produce complete 3D models. Recently, some works investigate 3D tasks with unconstrained images by interfering the 3D generation process, such as editing (Qi et al., 2024; Jin et al., 2025; Zheng et al., 2025), style-guided generation (Wang et al., 2024; Qu et al., 2025), morphing (Yin et al., 2025; Yang et al., 2025) and scene generation (Huang et al., 2025; Tan et al., 2021). However, these models are typically designed for specific tasks. To the best of our knowledge, no works have proposed a 3D generation pipeline that can process multiple unconstrained inputs of the same object.

3 Method

3.1 Overview

The overall pipeline of UMI3D is illustrated in Fig. 2(a). Given a set of unconstrained input images $\{I_k\}$ of the same object, UMI3D operates on a pre-trained 3D generative model that takes Gaussian noise as input and progressively denoises it into a 3D asset in a zero-shot manner under multi-image conditioning. The preliminary technique of our method is provided in the Appendix.

We first introduce a Voxel Reference Score (VRS) metric (Sec. 3.2), as shown in Fig. 2(b). Starting from the baseline cross-attention (cross-attn) between 3D voxel tokens and 2D image features, VRS analyzes the attention maps of each voxel v_i over different conditioning images and computes a per-image score that indicates how strongly the voxel is supported by each view. These VRS scores are then used to partition the 3D voxels into subsets preferentially associated with different input images. On top of VRS, we design the Simultaneous Focusing Cross-Attention (SFC-Attn) module (Sec. 3.3), as depicted in Fig. 2(c). All input images are encoded by a DINOv3 encoder into multi-view features $\{C_k\}$. Guided by the voxel-image assignments from VRS, SFC-Attn divides the 3D voxels into groups and performs cross-attn between each voxel group and its corresponding image features within the same DiT block. The outputs of different groups are then merged and propagated through the following network. UMI3D does not introduce additional trainable parameters; it only modifies the inference-time attention computation through VRS estimation and SFC-Attn routing.

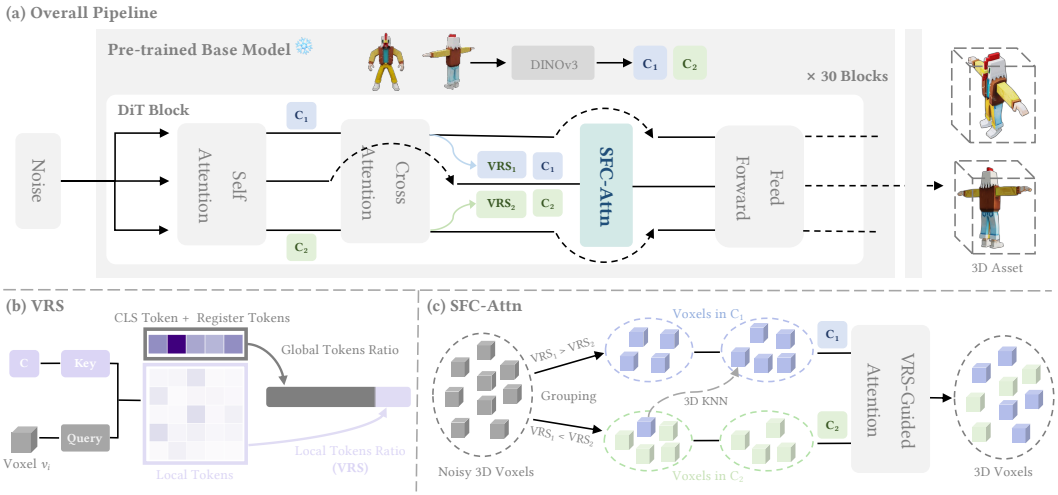

Figure 2: Overview of UMI3D. (a) Overall pipeline. (b) Voxel Reference Score (VRS). (c) Simultaneous Focusing Cross-Attention (SFC-Attn).

3.2 Voxel Reference Score (VRS)

For each conditioning image I_k, we extract features using a frozen DINO image encoder, obtaining a token sequence:

$$C_k = [\,\mathbf{t}^{\text{cls}},\, \mathbf{t}_1^{\text{reg}}, \ldots, \mathbf{t}_{N_r}^{\text{reg}},\, \mathbf{t}_1^{\text{loc}}, \ldots, \mathbf{t}_{N_l}^{\text{loc}}\,], \tag{1}$$

where $\mathbf{t}^{\text{cls}}$ is the classification token, $\{\mathbf{t}_{n_r}^{\text{reg}}\}_{n_r=1}^{N_r}$ are the register tokens, and $\{\mathbf{t}_{n_l}^{\text{loc}}\}_{n_l=1}^{N_l}$ are local tokens corresponding to image patches. We collectively refer to $\mathbf{t}^{\text{cls}}$ and $\mathbf{t}^{\text{reg}}$ as *global tokens*, $t^{\text{global}} = [\mathbf{t}^{\text{cls}}, \mathbf{t}^{\text{reg}}]$, as they capture rich global information about the entire image. In contrast, the local tokens are spatially aligned with image patches and primarily encode local appearances and fine-grained details.

By analyzing these cross-attn layers, we obtain two key insights that motivate our Voxel Reference Score (VRS): **(1) Global tokens receive disproportionately high attention per token.** As shown in Fig. 3 and Tab. 2, global tokens receive substantially higher average attention per token than local tokens, although their total attention mass varies across models and stages. This disproportionate allocation indicates that the backbone primarily consults global tokens to reason about the object's overall structure and appearance. **(2) Voxels lacking direct visual support fall back to global tokens.** Intuitively, a voxel that is well grounded in image I_k should rely more on the image's patch-aligned local tokens, which provide spatially specific evidence; conversely, a voxel without direct visual support in I_k has no localized evidence to attend to and instead falls back to the image-level semantics carried by global tokens. In the Fig. 4, when we replace the global tokens of the current view with those from another arbitrary image while keeping all local tokens fixed, the visible regions of the current view are only mildly affected, whereas the invisible regions change much more significantly. These observations suggest that global tokens implicitly encode how well each image explains different parts of the 3D space. This motivates us to explicitly quantify how strongly each voxel is supported by each conditioning image.

Specifically, for each voxel v_i and each input image I_k, we first obtain the corresponding query and key/value tokens from the cross-attn layer. The attention map of voxel v_i over the tokens of image I_k is then computed as $\text{AttnMap}_{i,k} = \text{softmax}\big(\frac{\mathbf{Q}_i \mathbf{K}_k^\top}{\sqrt{d_q}}\big)$, where d_q denotes the feature dimension. $\text{AttnMap}_{i,k} \in \mathbb{R}^{L_k}$ stores the attention weight from voxel v_i to each token of I_k. Based on $\text{AttnMap}_{i,k}$, we define the VRS value as the total attention assigned to the local tokens of image I_k, which is computed by subtracting the attention mass allocated to the global tokens:

$$\text{VRS}_{i,k} = \sum_{t \in t_{\text{local}}} \text{AttnMap}_{i,k}[t] = 1 - \sum_{t \in t_{\text{global}}} \text{AttnMap}_{i,k}[t]. \tag{2}$$

Here, $\text{VRS}_{i,k}$ measures how strongly voxel v_i is supported by the local evidence provided by image I_k – a higher value indicates that the voxel is well grounded in that image, while a lower value indi-

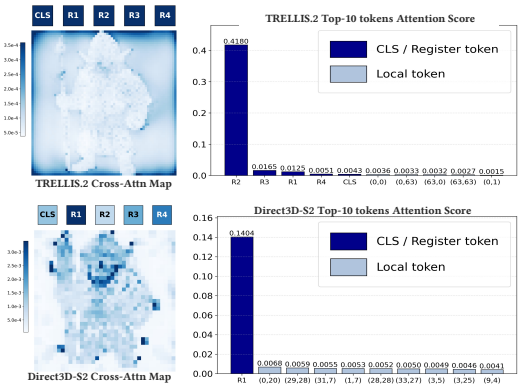

Figure 3: Global tokens receive disproportionately high attention per token.

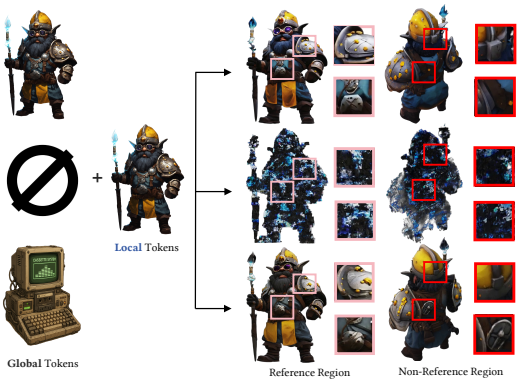

Figure 4: Replacing global tokens while keeping local tokens fixed affects non-reference regions more strongly than reference regions.

cates fallback to global semantics. By comparing $\mathrm{VRS}_{i,k}$ across different single-image conditioned generation branches k, we can determine which reference image provides the highest-confidence support for v_i. We refer the reader to Sec. 4.3 for empirical validation of the proposed VRS metric.

3.3 Simultaneous Focusing Cross-Attention (SFC-Attn) Module

We now describe how the proposed SFC-Attn module leverages the Voxel Reference Score (VRS) to enable simultaneous yet selective conditioning on multiple input images within each DiT block of the pre-trained 3D backbone, as illustrated in Fig. 2(a,c).

VRS Estimation Branches. Given the input images $\{I_k\}_{k=1}^{K}$ and their corresponding DINOv3 token sequences $\{C_k\}_{k=1}^{K}$, we first run multiple standard 3D conditioning branches, each conditioned on a *single* image. Concretely, for each image I_k, we feed its tokens C_k into the original cross-attn layer of the frozen base model and obtain the $\mathrm{VRS}_{i,k}$ for every voxel v_i, which measures how strongly voxel v_i is supported by image I_k. This stage consists of multiple *parallel* cross-attn passes, each identical to the original single-image conditioning of the base model, and serves only to estimate voxel-image affinities without modifying any backbone parameters. Although each branch produces branch-specific queries under its own conditioning image, the resulting VRS values remain comparable across branches: all branches share the same denoising step, DiT block, input voxel positions, and backbone parameters, differing only in the conditioning image. Therefore, VRS provides a consistent basis for comparing the relative support of different reference images for the same voxel. The validation experiments can be found in Sec. 4.3.

Voxel Grouping by VRS. For each voxel v_i, we define its dominant reference image as $\kappa(i) = \arg\max_k \mathrm{VRS}_{i,k}$, and construct K disjoint voxel groups:

$$\mathcal{G}_k = \{\, v_i \mid \kappa(i) = k \,\}, \quad k = 1, \ldots, K. \tag{3}$$

Thus, $\mathcal{G}_k$ contains the voxels whose 3D contents are most strongly supported by image I_k according to VRS. This produces a hard assignment that routes each voxel to a single conditioning image.

In 3D space, neighboring voxels typically belong to the same underlying surface region and therefore tend to be supported by the same conditioning image. To enforce such spatial coherence, we further refine the coarse VRS-based grouping with a 3D k-nearest-neighbor (KNN) scheme. Let the coordinate of voxel v_i be $\mathbf{p}_i = (x_i, y_i, z_i)$. For each voxel v_i, we consider its K_{nn} nearest neighbors in terms of 3D Euclidean distance $d(i, j)$, and denote their index set by $\mathcal{N}_K(i)$. We refine the routing decision of each voxel by taking the mode of the dominant views within its 3D KNN neighborhood:

$$\tilde{\kappa}(i) = \mathrm{Mode}\big(\{\kappa(j)\}_{j \in \mathcal{N}_K(i)}\big), \tag{4}$$

and update the voxel groups accordingly, as:

$$\tilde{\mathcal{G}}_k = \{\, v_i \mid \tilde{\kappa}(i) = k \,\}, \quad k = 1, \ldots, K. \tag{5}$$

This refinement brings two benefits: (i) it mitigates the influence of noisy or outlier VRS responses at individual voxels, leading to a more reliable view assignment; and (ii) it encourages ambiguous

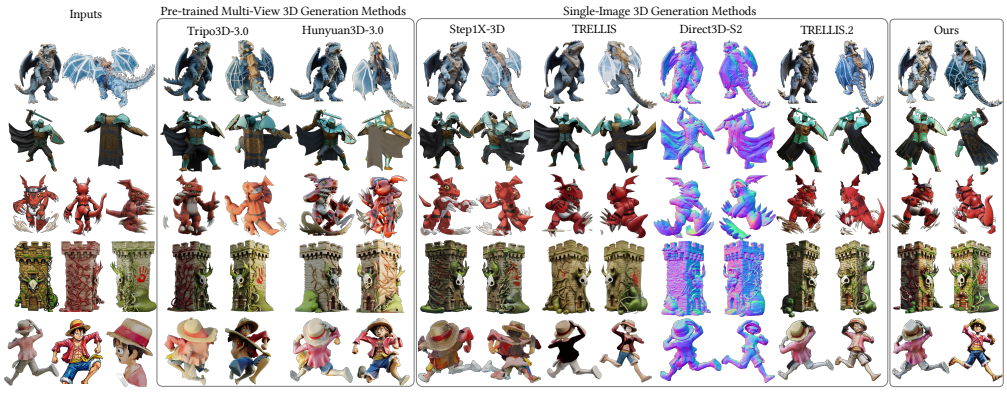

Figure 5: Qualitative comparison to multi-view 3D generation model.

voxels with weak direct evidence to follow the dominant reference of neighboring, better-supported voxels, thereby reducing feature discontinuities in ambiguous regions.

VRS-guided Attention. After obtaining the voxel groups $\{\tilde{\mathcal{G}}_k\}$, SFC-Attn performs view-specific cross-attn for each group. Let $\mathbf{Q}_k \in \mathbb{R}^{|\tilde{\mathcal{G}}_k| \times d}$ be the sub-matrix of queries corresponding to voxels in $\tilde{\mathcal{G}}_k$. For each image I_k, we compute keys $\mathbf{K}_k$ and values $\mathbf{V}_k$ from C_k, and then obtain the SFC-Attn output for voxels in $\tilde{\mathcal{G}}_k$ as $\mathbf{Z}_k = \mathrm{Softmax}\left(\frac{\mathbf{Q}_k \mathbf{K}_k^\top}{\sqrt{d_q}}\right)\mathbf{V}_k$, where $\mathbf{Z}_k \in \mathbb{R}^{|\tilde{\mathcal{G}}_k| \times d}$ contains the updated features of all voxels in group $\tilde{\mathcal{G}}_k$, extracted only from the conditioning image I_k. Finally, the group-wise outputs $\{\mathbf{Z}_k\}_{k=1}^K$ are scattered back to their original voxel indices to recover the full voxel feature matrix $\mathbf{Z}$. Since the refined groups $\{\tilde{\mathcal{G}}_k\}$ are disjoint by construction, this merge is unambiguous. In this way, all input images are *simultaneously* available within the same denoising step, while each voxel *focuses* on the single image that best explains it.

4 Experiments

In this section, we first conduct qualitative and quantitative comparisons, together with a user study, against existing state-of-the-art methods. We further perform ablation studies to validate the effectiveness of each proposed component, and validate the two insights underlying the VRS metric in Sec. 3.2. Finally, we showcase additional applications enabled by UMI3D. More results and implementation details can be found in **Appendix**.

4.1 Comparisons with State-of-the-Art Methods

4.1.1 Methods Compared.

To the best of our knowledge, UMI3D is the first approach that enables unconstrained multi-image input for 3D generation. Since no prior work directly addresses this task, we compare UMI3D with the approaches mentioned in Sec. 1: (1) pre-trained multi-view 3D generation approaches and (2) naive multi-image extensions of single-image 3D generation methods. For (1), we select Hunyuan3D 3.0 (Team, 2025) and Tripo3D 3.0 (Li et al., 2025b) for testing. Note that these are commercial models, and we use their official website interfaces for inference. For (2), we compare with Step1X-3D (Li et al., 2025a), TRELLIS (Xiang et al., 2024), Direct3D-S2 (Wu et al., 2025), and TRELLIS.2 (Xiang et al., 2025). For Direct3D-S2, TRELLIS and TRELLIS.2, only the anchor image is used as the condition in the sparse structure generation stage, following their default single-image settings. Since Direct3D-S2 cannot generate textures, we use the normal maps that it produces as a substitute in the visualizations.

4.1.2 Qualitative Results.

As shown in Fig. 5, we present extensive comparisons with existing SOTA 3D generation methods. (1) For *Pre-trained multi-view generation models*, Tripo3D and Hunyuan3D exhibit strong single-image 3D generation capabilities, but struggle to handle unconstrained multi-image inputs with

Table 1: **Quantitative evaluation.** All results are tallied across 40 different generation results.

Metric	CLIP ↑	DINO ↑	LPIPS ↓	User Study (%) ↑
Tripo3D-3.0 (Li et al., 2025b)	85.97	71.76	0.3361	4.98
Hunyuan3D-3.0 (Team, 2025)	85.41	69.67	0.3408	7.75
Step1X-3D (Li et al., 2025a)	85.52	69.01	0.3433	5.68
TRELLIS (Xiang et al., 2024)	85.60	70.44	0.3429	5.16
TRELLIS.2 (Xiang et al., 2025)	85.95	72.99	0.3413	15.86
UMI3D (Ours)	**86.42**	**74.62**	**0.3356**	**60.52**

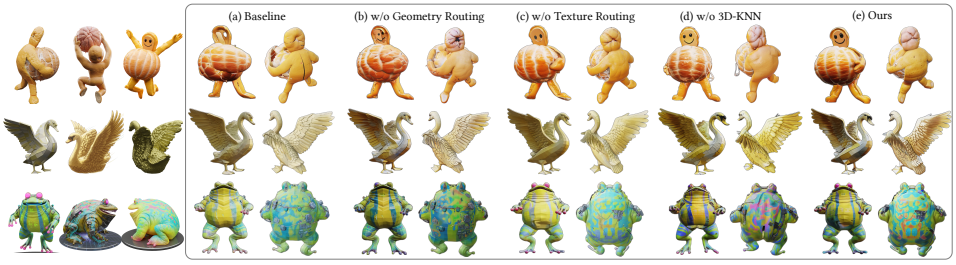

Figure 6: Overall ablation study on UMI3D. (a) The Baseline Model (TRELLIS.2). (b) w/o Geometry Routing. (c) w/o Texture Routing. (d) w/o 3D-KNN. (e) Our Full UMI3D Model.

heavy occlusions, large pose changes, or significant semantic differences. Since these models are trained with multi-view image sets that are strictly consistent in viewpoint and appearance, extending them to such unconstrained settings often leads to confused, misaligned, or entangled 3D results. (2) For *naive generation pipelines with single-image based models*, they are limited by the fact that only one image can be used as the condition at each step, which prevents the model from leveraging global information across all inputs and tends to cause problems like overfitting to a single condition, and missing, over-smooth or inconsistent geometry and textures. In contrast, UMI3D attends to all conditioning images simultaneously at each denoising step while selectively extracting the most informative evidence for each 3D region, yielding coherent and high-quality 3D generation under unconstrained multi-image inputs.

4.1.3 Quantitative Results.

The details of the metrics can be found in the **Appendix**. As shown in Tab. 1, our method achieves the best performance on all four metrics. The CLIP score mainly reflects high-level semantic consistency. Since all input images correspond to the same semantic object, the differences among the methods are relatively small. Nevertheless, UMI3D still outperforms the second-best method, Tripo3D 3.0, by 0.45 points. DINO is more sensitive to local details, and UMI3D improves the DINO score by 1.63 points over TRELLIS.2, achieving the best performance for unconstrained multi-image 3D generation. The user study provides the most discriminative signal: UMI3D receives **60.52%** of preference votes among the six evaluated methods, compared with 15.86% for TRELLIS.2, the runner-up, indicating that UMI3D produces better results for both reference and unseen viewpoints.

4.2 Ablation Study

To evaluate the effectiveness of each component, we conduct an ablation study as shown in Fig. 6. **Effectiveness of SFC-Attn in Geometry Generation.** As shown in (b) and (e), removing SFC-Attn from the geometry generation stage of UMI3D leads to errors in the structural information of the generated results (such as the face of the orange character), and inaccuracies in the structural information can also make it difficult to align the texture information, thereby affecting the results across multiple stages. **Effectiveness of SFC-Attn in Texture Generation.** The absence of SFC-Attn during the texture generation stage can lead to issues such as overly smooth textures (e.g., the color discrepancy at the front of the frog) and loss of detail (e.g., the eyes of the swan). **Effectiveness of the 3D-KNN operation.** Comparing (d) and (e), the absence of the 3D-KNN mechanism leads to some local discrepancies. Incorporating KNN processing improves routing accuracy, thereby producing high-quality 3D generation results.

4.3 INSIGHT VALIDATION

Here, we experimentally validate the two insights proposed in Sec. 3.2, providing empirical support for both the VRS metric and SFC-Attn.

Table 2: The attention score of different types of tokens.

Ratio (Per-Token Ratio)	CLS Token	Register Tokens	Local Tokens
TRELLIS (Stage-2)	0.0014 (0.0014)	0.8105 (0.2026)	0.1881 (0.0001)
Direct3D-S2 (Stage-2)	0.0022 (0.0022)	0.1541 (0.0385)	0.8437 (0.0006)
TRELLIS.2 (Stage-2)	0.0054 (0.0054)	0.4835 (0.1209)	0.5111 (0.0001)
TRELLIS.2 (Stage-3)	0.0050 (0.0050)	0.5335 (0.1334)	0.4615 (0.0001)

Table 3: Attention Routing Strategies.

Metric	CLIP ↑	DINO ↑	LPIPS ↓
Baseline	85.95	72.99	0.3413
Random CA Routing	85.61	72.36	0.3423
Opposite VRS Routing	84.42	69.63	0.3439
VRS Routing (Ours)	**86.42**	**74.62**	**0.3356**

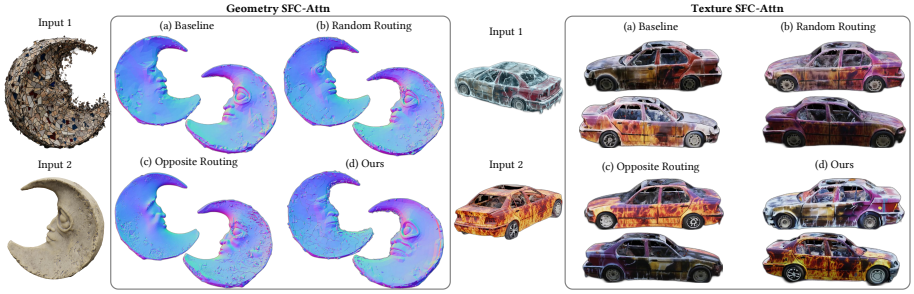

Figure 7: Insight validation of the SFC-Attn module.

To validate **Insight 1**, we analyze the attention scores assigned to different types of tokens in the condition image during single-image generation with different models. This analysis is conducted on 41 individual images from the official TRELLIS repository, which are separate from our 40-case multi-image benchmark. For each model, we average the attention scores over voxels and attention heads, and then accumulate them across all denoising steps. The resulting per-token scores, together with the number of tokens of each type, are reported in Tab. 2.

The results reveal that, when extracting information from a 2D image, these 3D generation models rely heavily on register tokens: in TRELLIS, the 4 register tokens alone account for 81% of the total attention, whereas the remaining 1,369 local tokens collectively contribute only about 19%. Although the CLS token receives a smaller share than the register tokens, its individual contribution is still substantially higher than that of any single local token. Based on this observation, we group the CLS token together with the register tokens as global tokens when computing the VRS metric.

To validate **Insight 2**, we present ablation results on UMI3D in Fig. 7 and Tab. 3. As shown in Fig. 7(a,b), for both geometry and texture generation, random routing without the VRS metric degrades the baseline performance, since it completely ignores the alignment between 3D voxels and the conditional images. Consistently, Tab. 3 shows that Random Cross-Attn Routing performs slightly worse than the baseline.

Fig. 7(c) further shows that selecting the image with the lowest VRS for each voxel introduces geometry and texture inconsistencies relative to the input images. Consistent with these qualitative observations, Tab. 3 shows that opposite-VRS routing reduces the DINO score from 72.99 to 69.63 compared with the baseline, while also worsening CLIP and LPIPS. In contrast, selecting the highest-VRS image improves all three metrics and produces more coherent results in Fig. 7(d). Together, these comparisons provide empirical support for VRS as a voxel-wise image-selection score, suggesting that higher-VRS images tend to provide more useful conditioning information for the corresponding voxels.

5 APPLICATIONS

Beyond unconstrained multi-image 3D generation, UMI3D also enables three practical applications: multi-view 3D generation enhancement, 3D morphing, and single-view local enhancement, as shown in Figs. 8, 9, and A.4. These examples further highlight the versatility of UMI3D. Detailed settings and analysis are deferred to the appendix.



Figure 8: The multi-view 3D generation enhancement results of UMI3D.



Figure 9: The 3D morphing results of UMI3D. α controls the 3D morphing progress.

6 Conclusion

This paper introduces *UMI3D*, a training-free multi-image 3D generation framework that supports robust 3D synthesis from unconstrained image collections. UMI3D incorporates a novel SFC-Attn module, which enables voxel-wise simultaneous and selective attention over multiple conditioning images to effectively integrate complementary visual information during generation. To address the ambiguity of voxel-image association in existing 3D generation models, UMI3D leverages the intrinsic behavior of cross-attention layers to derive the Voxel Reference Score (VRS), allowing the model to automatically establish reliable voxel-to-image correspondences without requiring external priors or auxiliary models. Finally, by reformulating multi-image conditioning as a dynamic routing process in the 3D latent space, UMI3D unlocks the multi-image capability of single-image 3D foundation models and achieves robust and coherent 3D generation in diverse real-world settings.

Limitations. As shown in Fig. A.5, UMI3D may fail to generate accurate 3D assets when the target object exhibits strong symmetry, such as a sphere or an object with multiple highly similar faces. This limitation arises because UMI3D relies on the implicit 2D-3D alignment priors embedded in the pre-trained 3D generative model to perform unconstrained generation. For such highly symmetrical objects, however, these implicit alignment cues become less reliable, which in turn degrades the effectiveness of VRS-based cross-attention routing and leads to unstable generation results.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing, code development and generating synthetic datasets. In particular, generative AI was used to improve the clarity, grammar of the manuscript, to assist in writing and debugging parts of the implementation code and to generate a validation benchmark of our proposed new task. All AI-assisted text was carefully reviewed and revised by the authors, and all AI-assisted code was manually tested and the authors verified the accuracy of the code. We did not use AI-generative tools to generate or formulate research ideas or core technical contributions of this work. We also did not use generative AI tools to search for, retrieve, or select references. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and code produced with the assistance of generative AI.

REFERENCES

Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2416–2425, 2023.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5939–5948, 2019.

Wei Cheng, Juncheng Mu, Xianfang Zeng, Xin Chen, Anqi Pang, Chi Zhang, Zhibin Wang, Bin Fu, Gang Yu, Ziwei Liu, et al. Mvpaint: Synchronized multi-view diffusion for painting anything 3d. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 585–594, 2025.

Xiang Feng, Chang Yu, Zoubin Bi, Yintong Shang, Feng Gao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, and Yin Yang. Arm: appearance reconstruction model for relightable 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21425–21437, 2025.

Junyao Gao, Jiaxing Li, Wenran Liu, Yanhong Zeng, Fei Shen, Kai Chen, Yanan Sun, and Cairong Zhao. Charactershot: Controllable and consistent 4d character animation. *arXiv preprint arXiv:2508.07409*, 2025.

Hyojun Go, Byeongjun Park, Jiho Jang, Jin-Young Kim, Soonwoo Kwon, and Changick Kim. Splatflow: Multi-view rectified flow model for 3d gaussian splatting synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21524–21536, 2025.

Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. Spaghetti: Editing implicit shapes through part aware generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–20, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23646–23657, 2025.

Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Ka-Hei Hui, Aditya Sanghi, Arianna Rampini, Kamal Rahimi Malekshan, Zhengzhe Liu, Hooman Shayani, and Chi-Wing Fu. Make-a-shape: a ten-million-scale 3d shape model. In *Forty-first International Conference on Machine Learning*, 2024.

Xuancheng Jin, Rengan Xie, Wenting Zheng, Rui Wang, Hujun Bao, and Yuchi Huo. Fuse3d: Generating 3d assets controlled by multi-image fusion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*, 2024.

Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv:2505.07747*, 2025a.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025b.

Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. *arXiv preprint arXiv:2505.14521*, 2025c.

Jiaqi Lin, Zhihao Li, Binxiao Huang, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Xiaofei Wu, Fenglong Song, and Wenming Yang. Decoupling appearance variations with 3d consistent features in gaussian splatting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5236–5244, 2025.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2837–2845, 2021.

Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7210–7219, 2021.

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4328–4338, 2023.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Zhangyang Qi, Yunhan Yang, Mengchen Zhang, Long Xing, Xiaoyang Wu, Tong Wu, Dahua Lin, Xihui Liu, Jiaqi Wang, and Hengshuang Zhao. Tailor3d: Customized 3d assets editing and generation with dual-side images. *arXiv preprint arXiv:2407.06191*, 2024.

Zefan Qu, Ke Xu, Gerhard Petrus Hancke, and Rynson WH Lau. Lush-nerf: Lighting up and sharpening nerfs for low-light scenes. *arXiv preprint arXiv:2411.06757*, 2024.

Zefan Qu, Zhenwei Wang, Haoyuan Wang, Ke Xu, Gerhard Petrus Hancke, and Rynson WH Lau. Stylesculptor: Zero-shot style-controllable 3d asset generation with texture-geometry dual guidance. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4209–4219, 2024.

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*, pp. 10684–10695, 2022.

Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David Fleet, and Andrea Tagliasacchi. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. *ACM Transactions on Graphics*, 44(2):1–11, 2025.

ByteDance Seed. Seed3d 1.0: From images to high-fidelity simulation-ready 3d assets. 2025.

J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20875–20886, 2023.

Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 892–901, 2021.

Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025.

Zhenwei Wang, Tengfei Wang, Gerhard Hancke, Ziwei Liu, and Rynson WH Lau. Themestation: generating theme-aware 3d assets from few exemplars. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, and Lu Sheng. Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21631–21641, 2025.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025.

Jiacong Xu, Yiqun Mei, and Vishal Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. *Advances in Neural Information Processing Systems*, 37:103334–103355, 2024.

Songlin Yang, Yushi Lan, Honghua Chen, and Xingang Pan. Textured 3d regenerative morphing with 3d diffusion prior. *arXiv preprint arXiv:2502.14316*, 2025.

Minghao Yin, Yukang Cao, and Kai Han. Wukong's 72 transformations: High-fidelity textured 3d morphing via flow models. *arXiv preprint arXiv:2511.22425*, 2025.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Yang Zheng, Hao Tan, Kai Zhang, Peng Wang, Leonidas Guibas, Gordon Wetzstein, and Wang Yifan. Splatpainter: Interactive authoring of 3d gaussians from 2d edits via test-time training. *arXiv preprint arXiv:2512.05354*, 2025.