# **A Technical Review and Analysis of a Ternary Diffusion Transformer for 3D Geometry Generation**

## **Section 1: Executive Summary and High-Level Verdict**

### **1.1. Overview of the Endeavor**

This report presents a comprehensive technical analysis of a custom-developed 3D geometry generation model. The model is based on the two-stage framework proposed in the Step1X-3D paper, which decouples geometry and texture synthesis for high-fidelity 3D asset creation.1 The work under review focuses exclusively on the first stage: geometry generation. A significant and ambitious architectural modification has been introduced. Instead of employing the Multi-Modal Diffusion Transformer (MMDiT) backbone specified in the reference paper, the implementation substitutes it with a Ternary Diffusion Transformer (TerDiT). This represents a novel research direction aimed at exploring the frontiers of model efficiency in 3D generation by leveraging extreme low-bit quantization, a technique primarily explored in 2D image synthesis and large language models.2 The objective of this report is to provide a rigorous evaluation of this architectural choice, analyze the fidelity of the implementation against the source paper, and offer a verdict on its viability, accompanied by a detailed roadmap for future development.

### **1.2. High-Level Verdict**

The decision to replace a full-precision MMDiT with a quantized TerDiT backbone is a theoretically compelling one. The primary motivation behind TerDiT is to drastically reduce the memory footprint and computational cost of large transformer models, with its original paper demonstrating over a 10-fold reduction in checkpoint size and an 8-fold reduction in inference memory for 2D image generation tasks.2 A successful application of this architecture to 3D latent space diffusion would be a significant contribution, potentially making large-scale, high-resolution 3D generative models more accessible for training and deployment in resource-constrained environments.

However, this ambitious modification introduces three fundamental and high-risk challenges that are not present in the original Step1X-3D framework and must be carefully managed:

1. **Representational Bottleneck:** A core conceptual conflict exists between the high-information, continuous latent space produced by the Step1X-3D Variational Autoencoder (VAE) and the low-precision, discrete nature of TerDiT's ternary weights. The VAE is meticulously designed to capture nuanced, high-frequency geometric details 1, while the ternarized linear layers of TerDiT have a vastly reduced capacity to represent complex, continuous functions. There is a substantial risk that the geometric detail encoded by the VAE will be "crushed" or aliased during processing by the TerDiT backbone, creating a severe information bottleneck that undermines the entire generation process.
2. **Training Instability:** The proposed training paradigm combines a Rectified Flow objective with Quantization-Aware Training (QAT). This is an untested and potentially unstable configuration. Rectified Flow learns a continuous velocity field to transport noise to data 1, while QAT involves a non-differentiable forward pass and relies on the Straight-Through Estimator (STE) for approximate gradients.4 The mismatch between the continuous nature of the learning objective and the discrete, error-prone nature of the model's computations and gradient estimation can lead to training divergence or poor convergence.
3. **Conditional Fidelity Loss:** The Step1X-3D model's ability to generate geometry from a single image is critically dependent on cross-attention mechanisms that fuse features from powerful image encoders (DINOv2, CLIP) with the 3D latent representation.1 Ternarizing the weight matrices within these cross-attention modules risks severely degrading the model's ability to process and align with these high-fidelity conditioning signals. The rich semantic and structural information from the image may be distorted or lost, resulting in generated shapes that poorly match the input condition.

### **1.3. Overall Recommendation and Report Structure**

The project is assessed as a theoretically plausible but high-risk research endeavor. Its success is not guaranteed and is highly contingent on a flawless data preparation pipeline and the careful, deliberate mitigation of the identified challenges related to information capacity, training stability, and conditioning fidelity. The choice of TerDiT is not an implementation error but rather an expert-level modification with profound and far-reaching consequences that must be actively managed.

This report is structured to serve as a detailed technical guide and roadmap for navigating these risks and maximizing the potential of this novel architecture. Section 2 provides a forensic analysis of the data curation pipeline, the non-negotiable foundation of the entire model. Section 3 performs a deep dive into the architectural implications of using TerDiT as a diffusion backbone for 3D latents, examining the critical interface points. Section 4 evaluates the training strategy, focusing on the potential instabilities of the chosen methodology. Finally, Section 5 synthesizes these findings into a conclusive verdict and provides a tiered set of actionable recommendations for remediation and future research.

## **Section 2: Analysis of the Data Curation and Preprocessing Pipeline**

### **2.1. Introduction: The Non-Negotiable Foundation**

The authors of the Step1X-3D paper explicitly state that "data scarcity constitutes the primary bottleneck in 3D generation advancement" and dedicate a significant portion of their report (Section 3.1) and a detailed diagram (Figure 3) to their rigorous data curation pipeline.1 This process, which refines over 5 million raw assets into a high-quality set of 2 million, is presented not as a preliminary step but as a core contribution and a key enabler of their state-of-the-art results.

For any generative model, the quality of the training data dictates the upper bound of the quality of its outputs. This principle is amplified to a critical degree when employing a quantized model like TerDiT. A full-precision model, with its 32-bit floating-point weights, possesses a vast parameter space and significant expressive capacity. This affords it a degree of flexibility to learn robust representations that can, to some extent, model and ignore noise, inconsistencies, or imperfections in the training data. A ternary model, by contrast, has its weights constrained to only three values: {−α,0,+α}.4 This dramatically reduces its expressive capacity. Consequently, a quantized model has significantly less "parametric room" to compensate for data imperfections. Any failure in the data pipeline—such as non-watertight meshes, incorrect surface normals, or poor alignment—will not be smoothed over by the model but will instead be amplified, leading to direct and visible artifacts in the generated geometry. Therefore, adherence to the Step1X-3D data curation protocol is not merely recommended; it is a non-negotiable prerequisite for the success of this project.

### **2.2. Component-wise Fidelity Assessment**

A systematic comparison between the data processing pipeline described in the user's script and the methodology detailed in Section 3.1 of the Step1X-3D paper is necessary. This assessment must verify fidelity at each stage.

#### **2.2.1. Data Filtering**

The Step1X-3D paper outlines a six-stage filtering process to eliminate low-quality assets. Each stage targets a specific failure mode common in web-sourced 3D data 1:

1. **Texture Quality Filtering:** This involves rendering six canonical views, converting them to HSV space, and filtering based on Hue and Value histograms to remove overly dark, bright, or uniform textures. A perceptual score is also computed to discard the bottom 20% of assets.
2. **Single-surface Filtering:** This uses Canonical Coordinate Maps (CCM) to detect and remove objects that are effectively 2D planes masquerading as 3D objects.
3. **Small Object Filtering:** This stage removes assets where the object of interest occupies less than 10% of the frontal view, ensuring the model trains on well-framed examples.
4. **Transparent Object Filtering:** This excludes assets with alpha-channel materials, as these cause a mismatch between rendered appearance and actual geometry, which is detrimental to training.
5. **Wrong Normal Filtering:** This crucial step identifies and removes meshes with inverted or incorrect normals by rendering normal maps and checking the angle between the normal vector and the camera position. Incorrect normals prevent successful watertight conversion.
6. **Name and Mesh Type Filtering:** A simple heuristic to remove assets explicitly labeled as point clouds, which are often noisy and difficult to process.

A review of the provided script must confirm that each of these six specific filtering steps is implemented as described. The absence of any one of these filters, particularly the wrong normal and transparent object filters, would allow corrupted data into the training set, severely impacting the model's ability to learn coherent geometric structures.

#### **2.2.2. Mesh-to-SDF Conversion**

A cornerstone of the Step1X-3D geometry pipeline is its use of a Truncated Signed Distance Function (TSDF) representation for geometric supervision. This requires all training meshes to be watertight. The paper highlights a critical improvement over prior methods: the "Enhanced mesh-to-SDF" process.1 Standard methods that rely solely on a "visibility check" (ray casting) can fail on complex meshes with holes or non-manifold geometry, creating internal "floaters" or incorrect surfaces. To solve this, Step1X-3D combines the visibility check with the calculation of a generalized winding number for points inside the mesh. This robustly determines inside/outside status even for imperfect geometry and is credited with a 20% improvement in the watertight conversion success rate.

The analysis of the user's script must verify whether this specific, enhanced technique is being used. If the implementation falls back to a simpler method (e.g., only visibility checks or a basic voxelization), it is highly probable that a significant portion of the dataset will have incorrect SDF representations, providing noisy and contradictory supervision signals to the VAE. For a quantized model, this noisy supervision would be catastrophic.

#### **2.2.3. Training Data Conversion**

The final stage of data preparation involves converting the clean, watertight meshes into the specific formats required by the VAE and the diffusion model.

* **Data for VAE:** The paper specifies two key strategies to help the VAE learn detailed geometry. First is **Sharp Edge Sampling (SES)**, where points are sampled not just uniformly from the surface (Puniform​) but also concentrated in areas of high curvature (Psalient​).1 This forces the VAE to pay more attention to fine details. Second is the use of three distinct sets of supervision points for the SDF decoder: points in the volume, points near the surface, and points on the surface. This provides a rich, multi-scale supervisory signal. The script must be checked for the presence of both SES and the multi-set SDF sampling.
* **Data for Diffusion:** To train the single-image-conditioned diffusion model, Step1X-3D renders each asset from 20 random viewpoints with specific camera parameters (elevation, azimuth, focal length) and applies augmentations.1 This process is vital for teaching the model view-invariance and ensuring it can generate a plausible 3D shape from any given 2D perspective. The script's rendering and augmentation protocol must match these specifications to ensure the diffusion model receives the same robust training distribution as the original.

### **2.3. Data Pipeline Fidelity Assessment Table**

The following table provides a structured summary for assessing the fidelity of the implemented data pipeline against the reference paper's specifications. This serves as a checklist to identify discrepancies and prioritize necessary corrections.

| Pipeline Stage | Step1X-3D Specification 1 | Implementation Status | Verdict & Recommendation |
| --- | --- | --- | --- |
| **Data Filtering** |  |  |  |
| Texture Quality | HSV histogram & perceptual score filtering. | *To be verified* | **Critical:** Must be implemented to remove aesthetically poor and uninformative assets. |
| Single-Surface | Canonical Coordinate Map (CCM) check. | *To be verified* | **Critical:** Prevents the model from learning degenerate 2D geometries. |
| Small Object | Frontal view alpha channel coverage > 10%. | *To be verified* | **Important:** Ensures the model trains on well-framed, clear examples. |
| Transparent Object | BSDF shader analysis to exclude alpha channels. | *To be verified* | **Critical:** Prevents mismatch between rendered images and underlying geometry. |
| Wrong Normals | View-space normal angle check. | *To be verified* | **Critical:** Essential for successful watertight conversion and correct SDF supervision. |
| Name/Mesh Type | Filter assets labeled as "point cloud". | *To be verified* | **Recommended:** A simple heuristic to remove a common source of noisy data. |
| **Mesh Conversion** |  |  |  |
| Mesh-to-SDF | Visibility check + Generalized Winding Number. | *To be verified* | **Critical:** The winding number is a key innovation for robust SDF generation. A simpler method is a major flaw. |
| **Data Conversion** |  |  |  |
| VAE Input Data | Sharp Edge Sampling (SES) for point clouds. | *To be verified* | **Critical:** SES is essential for preserving high-frequency geometric details in the VAE latent space. |
| VAE Supervision | Three distinct sets of SDF sampling points. | *To be verified* | **Important:** Provides a richer, more stable supervision signal for the VAE decoder. |
| Diffusion Input Data | 20 random views with specified camera params. | *To be verified* | **Critical:** Ensures the model learns a robust, view-invariant representation for image-to-3D generation. |

## **Section 3: Architectural Deep Dive: TerDiT as a Diffusion Backbone for 3D Latents**

### **3.1. Introduction: A Tale of Two Philosophies**

The architectural substitution at the heart of this project represents a fundamental clash of two distinct design philosophies. The original Step1X-3D model employs an **MMDiT (Multi-Modal Diffusion Transformer)** backbone, an architecture purpose-built for the high-fidelity fusion of multiple, full-precision data streams.1 As depicted in its architecture diagram (Figure 4), MMDiT uses a sophisticated hybrid of dual-stream and single-stream blocks. The dual-stream blocks process the 3D shape latent tokens and the image conditioning tokens in separate pathways, allowing them to retain their unique characteristics while interacting through cross-attention. The single-stream blocks then merge these tokens for joint processing, enabling deeper integration. This design prioritizes representational power and the precise alignment of multi-modal information.

In contrast, the **TerDiT (Ternary Diffusion Transformer)** architecture embodies a philosophy of extreme efficiency.2 Its primary goal is to minimize model size and memory usage to make large transformers deployable in resource-constrained settings. This is achieved through weight-only ternarization, a form of aggressive quantization where every weight in a linear layer is mapped to one of three values. The ternarization function can be expressed as:

f(W)=α⋅RoundClip(γ+ϵW​,−1,1)

where W is the full-precision weight matrix, γ is the mean absolute value of the weights in W, α is a learnable scaling factor for the entire layer, and ϵ is a small constant to prevent division by zero.4 During training, the non-differentiable rounding operation is handled by the Straight-Through Estimator (STE), which approximates the gradient by simply passing it through to the full-precision "shadow" weights during the backward pass. This philosophy prioritizes efficiency above all else, accepting a potential trade-off in expressive power based on the premise that large models possess significant precision redundancy.3 The core of this analysis lies in evaluating the consequences of replacing the fusion-oriented MMDiT with the efficiency-oriented TerDiT in the context of 3D latent diffusion.

### **3.2. The VAE-TerDiT Interface: A High-Risk Information Bottleneck**

The most significant theoretical risk introduced by this architectural substitution is the creation of a severe information bottleneck at the interface between the VAE and the diffusion backbone. The entire purpose of the Step1X-3D VAE, with its intricate design incorporating Sharp Edge Sampling and dual cross-attention, is to produce a rich, continuous, and high-dimensional latent vector set that faithfully captures the geometric essence of a 3D shape, including its fine details and high-frequency features.1 This latent set is not just a coarse representation; it is a carefully engineered information structure.

When this continuous latent vector is fed into the first layer of the TerDiT backbone, it is immediately subjected to a linear transformation defined by a matrix of ternary weights. A full-precision weight matrix can represent a nuanced, complex linear transformation. A ternary weight matrix, however, can only represent a very crude approximation of such a transformation. The function it computes is fundamentally coarse and piecewise-constant. The continuous, subtle variations in the VAE's latent vectors, which encode the difference between a sharp corner and a slightly rounded edge, are at high risk of being "crushed" or "aliased" into the same output value by the low-precision arithmetic of the ternary layer.

This process can be conceptualized as a lossy compression step being applied at every layer of the diffusion model. The geometric nuance that was so carefully preserved by the VAE may be progressively eroded with each pass through a TerDiT block. The model might learn the coarse, low-frequency aspects of the 3D shapes but struggle to reproduce the fine details that distinguish high-quality assets. This is the central information bottleneck: the expressive power of the diffusion backbone may be insufficient to fully leverage the richness of the latent space provided by the VAE, creating a ceiling on the achievable quality of the generated geometry.

### **3.3. Analysis of the TerDiT Implementation**

A code-level review of the implemented TerDiT architecture is crucial to ensure its correctness and stability. The original TerDiT research identified a critical failure mode when naively quantizing a standard Diffusion Transformer (DiT).4

The key finding was that the combination of the adaptive layer normalization (adaLN) module, which is conditioned on the diffusion timestep, and the ternary weights in the subsequent MLP layers led to excessively large activation values. These exploding activations destabilized the training process and prevented the model from converging. The solution proposed and validated in the TerDiT papers is a simple but non-obvious architectural modification: the introduction of an additional RMSNorm layer immediately following the MLP within the adaLN block. This normalization step effectively rescales the large activation values, stabilizing the training dynamics and enabling the successful training of multi-billion parameter ternary models.4

Therefore, a critical check of the implemented TerDiT block is to verify the presence of this specific RMSNorm layer. Its absence is not a minor deviation but a critical implementation flaw that would almost certainly replicate the training instability observed by the original TerDiT authors, making convergence highly unlikely. The model would likely suffer from exploding gradients and a loss that fails to decrease, regardless of hyperparameter tuning.

### **3.4. Efficacy of Conditional Signal Injection in a Quantized World**

The Step1X-3D model's capacity for single-image-to-3D generation hinges on its ability to effectively condition the diffusion process on image features. It achieves this by extracting powerful feature vectors from pre-trained encoders (DINOv2 and CLIP) and injecting them into the diffusion backbone via cross-attention mechanisms within the MMDiT blocks.1 In a cross-attention operation, the geometry latent (

x) is used to form the query (Q), while the image condition (c) is used to form the key (K) and value (V):

Q=xWq​,K=cWk​,V=cWv​Attention(Q,K,V)=softmax(dk​​QKT​)V

In the proposed TerDiT-based model, the projection matrices Wq​, Wk​, and Wv​ are ternarized. This poses a second major risk related to information fidelity. The rich, high-dimensional, continuous feature vectors from the DINOv2 and CLIP encoders, which capture intricate semantic and structural details of the input image, must be projected into the key and value spaces using these low-capacity ternary weight matrices.

This process is analogous to taking a high-resolution photograph and passing it through a filter that reduces its color depth to just three colors. The resulting K and V matrices will likely be a poor, distorted representation of the original conditioning signal. Consequently, the attention score computation (QKT) will be comparing a coarsely represented geometry query with a coarsely represented image key, leading to noisy, inaccurate, and potentially meaningless attention weights. The model may fail to learn the fine-grained alignment between specific parts of the image and specific parts of the geometry. The likely outcome is a significant degradation in conditioning quality. The model might learn to associate the image with the geometry in a very general sense (e.g., generating a "car-like" shape for an image of a car) but fail to replicate the specific pose, structure, or unique features of the car in the input image. This loss of conditioning fidelity would severely compromise one of the key capabilities of the Step1X-3D framework.

## **Section 4: Evaluation of the Training Strategy and Stability**

### **4.1. Introduction: Navigating Uncharted Training Dynamics**

The training strategy pairs a Rectified Flow objective with Quantization-Aware Training (QAT), a combination that ventures into uncharted territory in terms of training dynamics. The Step1X-3D paper validates the use of Rectified Flow for training its full-precision MMDiT model 1, and the TerDiT paper validates QAT for a standard denoising diffusion objective.2 The interaction between these two sophisticated techniques is novel and introduces a potential for significant instability that must be carefully analyzed and managed.

### **4.2. The Conflict Between Rectified Flow and QAT**

A fundamental conflict exists between the nature of the Rectified Flow learning target and the mechanics of QAT. The Rectified Flow objective is formulated to learn the velocity field of a linear probability path between noise and data. The loss function is:

L=Et,x0​,x1​​[∣∣uθ​(xt​,c,t)−ut​∣∣22​]

where ut​=x1​−x0​ is the ground-truth velocity, a continuous vector field, and uθ​ is the model's prediction of this velocity.1 The objective explicitly pushes the model to predict a smooth, continuous vector at every point in space-time.

However, the model predicting this velocity, uθ​, is a TerDiT. During the forward pass of QAT, its weights are ternarized, meaning its output is the result of a discrete, approximate computation. This introduces quantization error into the predicted velocity uθ​. More critically, the gradient of the loss with respect to the model's parameters is problematic. Since the ternarization function involves a non-differentiable rounding step, the true gradient is zero or undefined almost everywhere. QAT circumvents this by using the Straight-Through Estimator (STE), which, during the backward pass, approximates the gradient of the rounding function as if it were an identity function.4

This creates a "gradient mismatch." The loss is computed based on the error of the *quantized model's output* (how far the discrete prediction is from the continuous target), but the gradient used for updates is an approximation applied to the *full-precision shadow weights*. This mismatch can lead to unstable training dynamics. The gradient signal is derived from an objective that demands continuous precision, but it is being used to update shadow weights whose ultimate purpose is to produce a good *ternary* representation. The training process could easily become pathological, with the shadow weights chasing a continuous target that the ternary weights can never perfectly represent, leading to oscillations, stalled training, or divergence. The model is being asked to solve a continuous regression problem with a discrete tool, and the approximation used to bridge this gap (STE) may not be stable enough for the Rectified Flow objective.

### **4.3. Review of Optimizer and Hyperparameter Strategy**

The Step1X-3D paper describes a carefully designed two-phase training strategy to ensure stable convergence.1

* **Phase 1 (Rapid Convergence):** The model is first trained for 100k iterations with a smaller latent set size (512), a higher learning rate (1e-4), and a large batch size (1920). This phase allows the model to quickly learn the coarse features of the data distribution.
* **Phase 2 (Enhanced Precision):** The model is then trained for another 100k iterations with a larger latent set size (2048), a lower learning rate (5e-5), and a smaller batch size (960). This phase refines the model's capacity and precision on higher-dimensional data.

A review of the training script must verify if this two-phase strategy is adopted. For a QAT process, which is inherently less stable than full-precision training, this strategy is not just beneficial but likely essential. The initial, volatile phase of QAT, where the shadow weights are learning to align with their ternary counterparts, would benefit immensely from the simpler problem setting of the smaller latent dimension in Phase 1. Attempting to start QAT directly with the large 2048-dimensional latent space and a high learning rate is a recipe for instability.

Furthermore, QAT often requires different hyperparameter considerations than standard training. A lower learning rate and a more gradual learning rate warmup schedule are typically necessary to prevent the approximate gradients from causing large, destabilizing updates to the shadow weights. Careful monitoring of the training process is paramount. This includes logging and visualizing the distributions of activations and the norms of gradients for both the main latent processing stream and the conditional stream. Any signs of explosion or vanishing in these values would be an early indicator of the instability predicted by the analysis and would require immediate intervention, such as further reducing the learning rate or adjusting weight decay.

## **Section 5: Synthesis, Final Verdict, and Actionable Recommendations**

### **5.1. Conclusive Verdict: A High-Risk, High-Reward Research Direction**

The synthesis of the preceding analysis leads to a nuanced verdict. The endeavor to replace the MMDiT backbone of the Step1X-3D geometry generator with a quantized TerDiT architecture is a forward-thinking and intellectually ambitious project. If successful, it would yield a 3D generative model with a dramatically reduced memory and computational footprint, a significant practical contribution to the field.

However, the project, as conceived, faces severe theoretical and practical hurdles that make success uncertain without careful and deliberate intervention. The analysis has identified three primary points of high risk:

1. **Information Bottleneck:** The limited expressive power of the ternary weights is fundamentally at odds with the high-fidelity, continuous latent space produced by the VAE, risking the loss of essential geometric detail.
2. **Conditioning Fidelity Loss:** The ternarization of cross-attention projection matrices is likely to corrupt the high-information image conditioning signal, severely weakening the model's image-to-3D generation capability.
3. **Training Instability:** The novel combination of a continuous Rectified Flow objective with the discrete and approximate nature of Quantization-Aware Training creates a high potential for unstable training dynamics and divergence.

Therefore, the final verdict is that the model is unlikely to match the performance of the original Step1X-3D framework without substantial modifications to mitigate these specific issues. The choice of TerDiT should not be viewed as an error, but rather as an expert-level modification that introduces a cascade of complex challenges. The path forward requires moving from a direct replacement strategy to a more nuanced co-design strategy, where the architecture and training process are adapted to accommodate the constraints of extreme quantization.

### **5.2. Tiered Recommendations for Improvement**

The following prioritized, actionable recommendations are provided to guide future work on this project. They are structured in tiers, from critical fixes required for basic stability to more advanced research directions for maximizing performance.

#### **5.2.1. Tier 1: Critical Fixes for Stability and Foundational Integrity**

These recommendations address fundamental flaws that will almost certainly prevent the model from training successfully. They are prerequisites for any further experimentation.

1. **Ensure Data Pipeline Parity:** The first and most critical action is to conduct a rigorous audit of the data processing script and ensure 100% compliance with the Step1X-3D data pipeline as detailed in Section 3.1 of the paper.1 This includes all six filtering stages, the use of the  
   **generalized winding number** for robust mesh-to-SDF conversion, and the implementation of **Sharp Edge Sampling (SES)** for VAE data preparation. This is a non-negotiable foundation.
2. **Verify and Implement Critical RMSNorm Layer:** The TerDiT architecture is known to be unstable without an additional RMSNorm layer in its transformer blocks to control activation magnitudes.4 The implementation must be immediately checked for this component. If it is missing, it must be added to prevent the near-certainty of activation explosion and training failure.
3. **Implement Robust Training Monitoring:** Implement comprehensive logging and visualization using tools like TensorBoard or Weights & Biases. Specifically, track the distributions (histograms) and norms of activations and gradients over time. This should be done separately for the latent processing stream and the conditional processing stream to provide early warnings of instability, allowing for proactive hyperparameter adjustments.

#### **5.2.2. Tier 2: Improving Performance and Information Flow**

Once the model is stable, these recommendations aim to directly address the information bottleneck and conditioning fidelity issues to improve generative quality.

1. **Adopt a Hybrid-Precision Architecture:** Instead of quantizing the entire model, adopt a hybrid-precision approach. The most vulnerable component is the cross-attention interface for the image condition. As a primary mitigation strategy, the cross-attention projection matrices (Wk​ and Wv​ that process the DINOv2/CLIP features) should be kept in full (FP32) or half (BF16) precision. This creates a "full-precision interface" for the conditioning signal, allowing the rich image information to enter the model without being distorted by ternarization, directly mitigating the fidelity loss identified in the analysis. The self-attention and MLP layers that process the geometry latent could remain quantized.
2. **Explore VAE-Quantization Co-design:** For a more advanced solution to the information bottleneck, consider fine-tuning the pre-trained VAE with a quantization-aware objective. This would involve adding a new term to the VAE's loss function that encourages it to produce latent vectors that are more "quantization-friendly." For example, a clustering loss could push the latent vector components towards values that are more likely to survive the ternarization process in the subsequent TerDiT layers. This co-design approach would proactively shape the latent space to be more compatible with the limited capacity of the diffusion backbone.
3. **Strictly Adhere to the Two-Phase Training Schedule:** The two-phase training strategy from the Step1X-3D paper is critical for stabilizing QAT.1 Begin training with a smaller latent dimension (512) to allow the model to learn the basics of the QAT process in a more stable regime before scaling up to the full 2048-dimensional latent space for fine-tuning.

#### **5.2.3. Tier 3: Advanced Research and Ablation**

These recommendations are for longer-term research to fully understand the trade-offs and explore the design space of efficient 3D generative models.

1. **Ablate Quantization Strategy:** Ternary quantization may be too aggressive for this task. It is recommended to explore less extreme but still highly efficient quantization schemes, such as 4-bit or 8-bit integer quantization (INT4/INT8). These methods offer a better balance between efficiency and expressive power and may prove to be a more suitable middle ground for this high-fidelity generation task.
2. **Conduct a Formal Ablation Study:** To precisely identify where performance is being lost, design a structured set of experiments to ablate the effect of quantization on different model components. Train separate models where only the self-attention layers are quantized, only the MLP layers are quantized, and only the cross-attention layers are quantized. Comparing the performance of these models against a fully quantized and a full-precision baseline would provide invaluable, publishable insights into which parts of the transformer architecture are most sensitive to quantization in the context of 3D latent diffusion.
3. **Investigate Alternative Conditioning Mechanisms:** If hybrid-precision cross-attention still proves insufficient or unstable, an alternative is to explore simpler conditioning mechanisms that might be more robust to quantization. For example, instead of cross-attention, the image conditioning signal could be used to predict the scale and shift parameters in an adaptive layer normalization (adaLN) scheme. This removes the problematic matrix multiplications on the conditioning vector and might offer a more stable, albeit potentially less powerful, way to inject the conditional information.

#### Works cited

1. Text\_to\_3d(mesh and texture).pdf
2. TERDIT: TERNARY DIFFUSION MODELS WITH TRANS- FORMERS - OpenReview, accessed July 12, 2025, <https://openreview.net/pdf/ed79b6f6c442a19f14ec876de59e714482aaacc9.pdf>
3. TerDiT: Ternary Diffusion Models with Transformers - arXiv, accessed July 12, 2025, <https://arxiv.org/html/2405.14854v2>
4. [Revue de papier] TerDiT: Ternary Diffusion Models with Transformers, accessed July 12, 2025, <https://www.themoonlight.io/fr/review/terdit-ternary-diffusion-models-with-transformers>
5. TerDiT: Ternary Diffusion Models with Transformers | OpenReview, accessed July 12, 2025, <https://openreview.net/forum?id=MVQj0uajaF>