

Beyond Prompts: Unconditional 3D Inversion for Out-of-Distribution Shapes

Victoria Yue Chen¹, Emery Pierson², Léopold Maillard^{2,3}, and
Maks Ovsjanikov²

¹ ETH Zürich

² LIX, École Polytechnique, IP Paris

³ Dassault Systèmes

Abstract. Text-driven inversion of generative models is a core paradigm for manipulating 2D or 3D content, unlocking numerous applications such as text-based editing, style transfer, or inverse problems. However, it relies on the assumption that generative models remain sensitive to natural language prompts. We demonstrate that for state-of-the-art native text-to-3D generative models, this assumption often collapses. We identify a critical failure mode where generation trajectories are drawn into latent “sink traps”: regions where the model becomes insensitive to prompt modifications. In these regimes, changes to the input text fail to alter internal representations in a way that alters the output geometry. Crucially, we observe that this is not a limitation of the model’s *geometric* expressivity; the same generative models possess the ability to produce a vast diversity of shapes but, as we demonstrate, become insensitive to out-of-distribution *text* guidance. We investigate this behavior by analyzing the sampling trajectories of the generative model, and find that complex geometries can still be represented and produced by leveraging the model’s unconditional generative prior. This leads to a more robust framework for text-based 3D shape editing that bypasses latent sinks by decoupling a model’s geometric representation power from its linguistic sensitivity. Our approach addresses the limitations of current 3D pipelines and enables high-fidelity semantic manipulation of out-of-distribution 3D shapes. Project webpage: <https://daidedou.sorpi.fr/publication/beyondprompts>

Keywords: Text-to-3D Editing · Out-of-Distribution Inversion · 3D Generative Models

1 Introduction

Inverting generative models consists of mapping a given input into the noise space of a pre-trained network. In recent years, this approach has been shown to enable many analysis and editing applications thanks to its flexibility and efficiency (*e.g.*, [29, 41] among myriad others). The process typically relies on a paired visual input (*e.g.*, an image or a 3D shape) and an approximate text prompt, representing this input. By embedding the source (*e.g.*, “rabbit”) in the

model noise space, text-based editing and manipulation can be done by simply modifying the prompt (*e.g.*, “rabbit lunging forward”), followed by denoising to obtain the edited asset. This simplicity, where the user only has to modify a text prompt instead of relying on complex, specialized software, can be leveraged in many scenarios such as text-driven editing [29], style transfer [49], or image inverse problems [3].

Despite the appeal of this strategy, we observe that current 3D manipulation pipelines do not *exclusively* use a 3D generative model or follow the simple strategy described above. Instead, to achieve plausible results, they often require auxiliary guidance such as user-defined masks, 2D in-painting models, or exploit strong image-based priors [21, 31, 46]. *Exclusively* using text prompts has not yet been shown to output high-quality inversion or editing results of native text-to-3D generative models.

In this paper we analyze the root causes behind this phenomenon and reveal a *fundamental mismatch* between the *geometric* expressivity of text-to-3D generative models and their *text*-reasoning capacity. Specifically, we note that, as illustrated in Fig. 1, edit failures often do not stem from the generative model’s inability to represent a target *geometry*, but rather from the prompt falling out-of-distribution. We identify this failure mode as the “sink trap”: a state where the model becomes unresponsive to textual variations, consistently producing a nearly identical shape regardless of the prompt, despite possessing the capacity to produce a variety of geometric structures.

We analyze this behavior by investigating the internal sampling trajectories of 3D generative models, and observe that they become highly unstable when using mismatched or out-of-distribution prompts during inversion. Based on this analysis, we find that leveraging the *unconditional distribution* of the generative model is crucial to producing stable trajectories. Building on these insights, we propose an approach that enables the faithful inversion of complex, non-rigid shapes that fail with the standard null-text inversion with approximate prompts [29] in the context of text-to-3D generative models. Remarkably, despite using the unconditional prior during inversion, we show that our approach still enables *3D shape editing*, driven purely by the target text prompt without the need for any auxiliary models or image-based priors. To the best of our knowledge, ours is the first approach that demonstrates high-fidelity 3D editing using *only a* native text-to-3D generative model and without any additional input beyond the target prompt. Our contributions are summarized as follows:

1. Characterization of the “sink trap” phenomenon: We identify and analyze a specific failure mode in 3D generative models where textual conditioning fails to trigger geometric changes, showing that these failures often stem from a mismatch between a model’s text reasoning and geometric representational capabilities.
2. Trajectory Stabilization via Unconditional Priors: We provide a study of the internal sampling trajectories of 3D models and introduce a method to stabilize them by leveraging the unconditional distribution. This enables the re-

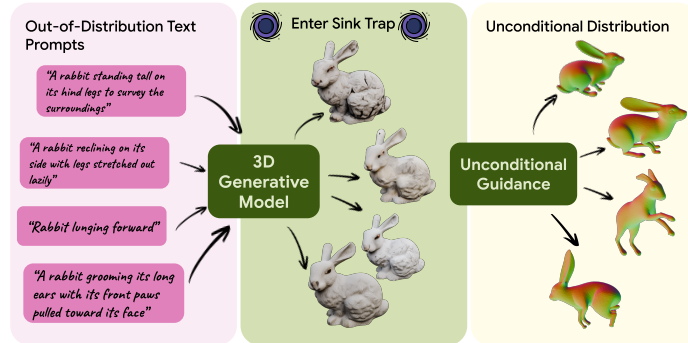


Fig. 1: Geometry vs. Language Diversity. (Left) Text-conditioned generation exhibits a “sink trap” effect, where diverse prompts for a “rabbit” yield nearly identical geometries. (Right) In contrast, our unconditioned 3D generative model overcomes this linguistic bottleneck, faithfully inverting and reconstructing arbitrary 3D shapes with high fidelity.

construction of complex, non-rigid shapes where standard text-conditioning cannot succeed.

3. A Text-to-3D Editing Framework: We show that our proposed inversion unlocks the first high-fidelity 3D editing pipeline driven solely by a target text prompt. Our approach relies solely on a native 3D generative model, removing the need for auxiliary image-based priors.

2 Related Work

2.1 Image Based Inversion

Inversion is a prerequisite for editing real-world images with Text-to-Image (T2I) models, as it maps a data sample back to a latent code within the model’s noise space by reversing the sampling process. This conversion provides a representation that the model can manipulate, a process fundamental to text-driven editing with diffusion models [9, 37, 38]. However, direct DDIM inversion often yields imperfect reconstructions due to numerical error accumulation [29, 40] and the challenges of classifier-free guidance (CFG) [10]. While various methods address these inaccuracies (see [11] for a survey), Null-Text Inversion (NTI) [29] remains a cornerstone approach. By optimizing unconditional embeddings, NTI ensures that the edited result maintains high fidelity to the original image while remaining responsive to new textual prompts.

Similar principles have recently been extended to flow-based generative models [24, 26], including *inversion-free* frameworks [12, 16, 44] that bypass iterative optimization and mapping to the noise space. Notably, these inversion-free methods still require mapping the target into a latent representation and necessitate a

source prompt that describes the original content. As observed in NTI [29], any prompt that is “reasonable and editable” would work. This requirement persists when extending 2D solutions to 3D: both the source shape and a corresponding text prompt are required to anchor the asset within the model’s latent space for subsequent inversion or editing. Despite the prominence of these methods in the image domain, the applicability of these latent-space manipulation techniques to the native latent spaces of 3D generative models remains largely unexamined.

2.2 3D content manipulation with Generative Models

Common approaches to 3D asset manipulation can be categorized into three paradigms: (1) Score Distillation Sampling (SDS) [32], which optimizes assets via image generative priors; (2) multiview-consistent image generation, which leverages multiview priors to ensure geometric coherence; and (3) direct manipulation of 3D representations coupled with 2D inpainting models.

Early work in 3D editing primarily relied on distilling gradients from pre-trained 2D diffusion models [4, 14, 36]. By rendering 3D assets from stochastic viewpoints, a generative model can output a score that guides optimization toward a text-aligned result. While applicable to both Neural Radiance Fields (NeRF) and explicit meshes, this per-edit optimization process is computationally expensive and slow.

Consequently, many subsequent works leverage multiview image generation for 3D editing [2, 5, 6, 22, 27]. These methods are generally more efficient, relying on multiview-consistent diffusion models to generate edited views, followed by a rapid reconstruction or refinement step to ensure global geometric consistency.

Finally, emerging works attempt to leverage state-of-the-art 3D models [43] directly for text-based editing [21, 46]. Notably, [46] employs Gemini 2.5 Flash [7] for inpainting, feeding the edited result back into [43] for asset generation, while VoxHammer [21] utilizes Flux [17] for inpainting rendered assets. Rather than manipulating internal model latents, these works introduce the editing signal by modifying the conditioning input. Their primary contribution lies in leveraging internal voxel representations to ensure the edited asset maintains structural fidelity to the original. More recently, [31] used inversion in the multi-view space. However, due to instabilities in the inversion process, they rely heavily on 2D backbones for mask selection and multi-view optimization of the geometry.

In summary, existing 3D editing frameworks rarely rely exclusively on the native generative capabilities of 3D models, instead incorporating 2D text-to-image (T2I) priors to ensure high visual fidelity. Our study isolates the 3D generative model to investigate the inherent properties of its latent space.

3 Preliminaries

3.1 Problem formulation

Our goal is to invert a 3D mesh in the latent space of a text-to-3D flow model. Namely, our inputs are: (1) a 3D shape $\mathcal{X}$ embedded in the representation space

of the flow model as a 3D-VAE latent z_0 , (2) the conditioning signal C paired with the source shape (which takes the form of a text prompt $\mathcal{P}_{src}$), and (3) a trained text-to-3D flow model $v_\theta(x, t, C)$ that can generate 3D shape latents z from noise.

Our output is a noisy latent z_1 , such that when re-integrating the generation process, the generated $\hat{z}_0$ matches z_0 closely. This inversion process is necessary as we want the mesh to be in a representation that the generative model can manipulate. When the process of inversion is successful, editing becomes a seamless operation: we can simply use an edit prompt $\mathcal{P}_{edit}$ and $\hat{z}_0$ to obtain the desired shape.

3.2 Rectified Flow and Inversion

Rectified flow models [24, 26] learn a velocity field $v_\theta(x_t, t)$ that transports a simple prior p_1 to a target data distribution p_0 by integrating the ordinary differential equation (ODE) $dx_t/dt = v_t(x_t)$. Sampling integrates the ODE from $t = 1$ to $t = 0$. Conversely, inversion maps a data latent z_0 back to noise z_1 by integrating the ODE forward ($t : 0 \rightarrow 1$). Following a deterministic baseline analogous to DDIM inversion [37], we utilize forward Euler steps:

$$z_{t_{i+1}} = z_{t_i} + (t_{i+1} - t_i)\hat{v}_\theta(z_{t_i}, t_i, C), \quad (1)$$

where $\hat{v}_\theta$ is the velocity field modified by classifier-free guidance (CFG) [10]. Specifically, given a conditioning signal C and guidance scale w , the guided velocity is computed as $\hat{v}_\theta(z_t, t, C) = v_\theta(z_t, t, \emptyset) + w(v_\theta(z_t, t, C) - v_\theta(z_t, t, \emptyset))$, where $\emptyset$ represents the null condition.

Inversion is the reverse process of generation: mapping any given input z to the noise space. Given a data latent z_0 , one can invert the rectified flow to recover the corresponding noise latent z_1 by integrating the ODE in the forward direction using Eq. 1 with $t_0 = 0$ and $t_N = 1$ [18, 35]. We call this process Euler inversion.

Null-Text Inversion (NTI) [29] improves reconstruction quality when text-shape alignment is imperfect. Since direct Euler inversion can result in inaccurate reconstructions, NTI optimizes the unconditional embeddings along the forward trajectory. During sampling, these optimized null embeddings ensure the generation trajectory closely matches the inverted reference path, preserving the original shape’s identity while enabling subsequent prompt-based editing.

$$v_\theta(x_t, t) = (1 + \omega) v_\theta(x_t, t, C) - \omega v_\theta(x_t, t, \emptyset), \quad (2)$$

3.3 3D Generative Backbone

Our study utilizes TRELLIS [43], a state-of-the-art open-source framework for high-fidelity *text-to-3D* generation. We do not consider recent iterations of the

method, including TRELLIS.2 [42], as its improvements primarily target the image-to-3D branch, which falls outside the scope of our study. TRELLIS generates assets in two stages:

Sparse Structure Generation. A first rectified flow model $\mathcal{G}_S$ produces a low-dimensional latent $z_0 \in \mathbb{R}^{8 \times 16 \times 16 \times 16}$, which is decoded into a coarse 64^3 voxel-based structure. This stage defines the global geometry of the asset. Any mesh can be projected into this space by voxelization and encoding via the pre-trained VAE (more details in Supp. Sec. B).

Structured Latent Generation. A second model $\mathcal{G}_L$ refines the coarse structure by generating high-dimensional sparse latents (SLATs), which are necessary for creating the resulting texture details. These latents are then decoded into the final 3D representations, such as meshes, radiance fields or Gaussian splats. Our study primarily operates within the latent space of the structure model $\mathcal{G}_S$. Modifying $\mathcal{G}_L$ predominantly results in texture-level updates with only marginal changes to the underlying geometry (see Supp. Sec. B).

4 Expressivity of Language and Geometry

Our key observation is that the assumption that textual instructions can reliably navigate the latent space of 3D models often fails when prompts fall out-of-distribution. We first quantify this failure by measuring the model’s responsiveness to different semantic categories and identifying where generation collapses into “sink traps”. We then evaluate how different text prompts impact inversion stability compared to image-based models. Finally, we demonstrate that by bypassing text-conditioning, the underlying generative prior can be used to create and edit complex shapes.

4.1 Language Diversity

The generative fidelity of 3D generative models is non-uniform across the prompt distribution. We observe that for specific semantic categories, the model exhibits a form of localized model collapse. In these regions, the model loses sensitivity to textual variations and converges toward a generic shape. This is qualitatively shown in Fig. 2, where diverse prompts regarding specific subjects yield near-identical topology. Note the variation across semantic categories.

To quantify this behaviour, we perform a diversity analysis using SigLIP [39]. We generate a collection of prompts $\mathcal{T} = \{t_1, \dots, t_n\}$ centered on specific semantic anchors (*e.g.*, astronaut, labrador) using Gemini 3 Pro [8] and create associated shapes with TRELLIS [43].

For each generated asset, we obtain a set of multi-view renders $\mathcal{I}_i$ covering a 360° orbit. Let $\Phi_{\text{vis}}(\cdot)$ and $\Phi_{\text{txt}}(\cdot)$ denote the SigLIP image and text encoders,



Fig. 2: We generate multiple assets using TRELLIS [43] across diverse character classes (e.g., surgeon, astronaut) while varying specific prompt attributes (e.g., “[Class] in a sitting pose” vs. “sprinting with arms swinging back and forth”). Despite these targeted variations, we observe significant mode collapse, where the model converges to a nearly identical geometry and texture for some classes, failing to modify the final result as requested by the text prompt.

respectively. We define the average pairwise distance for both modalities as:

$$\begin{aligned}\Delta_{\text{vis}} &= \frac{1}{N(N-1)} \sum_{i \neq j} \left(1 - \cos(\Phi_{\text{vis}}(\mathcal{I}_i), \Phi_{\text{vis}}(\mathcal{I}_j)) \right) \\ \Delta_{\text{txt}} &= \frac{1}{N(N-1)} \sum_{i \neq j} \left(1 - \cos(\Phi_{\text{txt}}(t_i), \Phi_{\text{txt}}(t_j)) \right)\end{aligned}\quad (3)$$

The diversity ratio $R = \Delta_{\text{vis}}/\Delta_{\text{txt}}$ characterizes the model’s responsiveness to linguistic shifts. This ratio is closely related to the Lipschitz constant of the generative model with respect to text inputs, where the main difference is that the cosine distance does not necessarily define a metric space.

A low ratio ($R \ll 1$) indicates that the model is operating in a sparse region of the joint embedding space, where noticeable changes in the text prompt lead to small (or no) visual changes. This behavior can be attributed to the text capability of the generative model being effectively out-of-distribution (OOD) relative to the learned geometric priors. In such cases, the conditional signal fails to steer the generation, causing the output to collapse toward a generic canonical shape. Our quantitative

Table 1: Values of Δ_{vis} , Δ_{txt} , and R across different subjects for TRELLIS [43], LN3Diff [19] (LN), and GaussianAnything [20] (GA). Most subjects exhibit $R < 1$.

Subject	TRELLIS			LN	GA
	Δ_{vis}	Δ_{txt}	R	$\frac{R}{R}$	$\frac{R}{R}$
Cute puppy	0.376	0.394	0.990	0.970	0.538
Surgeon	0.196	0.377	0.515	0.806	0.613
Husky	0.381	0.427	0.799	0.651	0.496
Dancing girl	0.512	0.404	1.205	0.626	0.569
Car	0.420	0.367	1.124	0.417	0.452
Scary wolf	0.156	0.270	0.640	1.041	0.707
Astronaut	0.260	0.461	0.638	0.861	0.529
Labrador	0.441	0.489	0.788	0.697	0.389

results in Tab. 1 suggest that the text-to-shape manifold is non-uniformly populated. Some specific semantic groups reside in densely mapped regions that facilitate high diversity, while others encounter sink regions with little to zero diversity. To confirm this phenomenon generalizes beyond TRELLIS, we additionally evaluated LN3Diff [19] and GaussianAnything [20] models. As shown in Tab. 1, both models consistently exhibit $R < 1$ across the majority of categories. We also evaluated models prior to 2024 [13, 33, 47], however their quantitative performance was too poor to generate plausible results for many input prompts.

This poses a significant challenge for 3D shape editing: in general, the user does not have access to the “ground truth prompt” of the shape, but rather an approximate prompt based on their observation, or the help of an auxiliary tool (*e.g.*, LLM). Because of the existence of sparse regions in the joint embedding space, the approximate prompt will likely be out-of-distribution (OOD). Therefore, using an approximate prompt will unlikely yield the desired result as the model will become unresponsive to it.

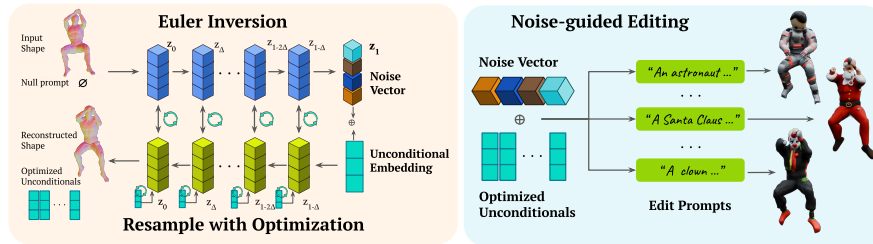


Fig. 3: Our unconditional 3D shape inversion (Left) and text-driven editing (Right) pipelines. We invert an arbitrary input shape by using an empty prompt and refining its embedding via NTI optimization [29]. Remarkably, this unconditional inversion strategy not only yields superior reconstructions, but the resulting noisy latent and optimized embedding also supports powerful, open-vocabulary editing of the input 3D shape.

4.2 Out-of-Distribution Prompts

As inversion requires a source prompt that corresponds to the initial geometry, when the user provides an OOD prompt relative to the shape, the joint embedding fails to align, and the resulting reconstruction will fail. Worse, any downstream task, such as editing, will be affected, and the output will not correspond to the user’s intent. This behaviour differs from T2I models, where it has been noted that providing “any reasonable and editable prompt” is sufficient for robust inversion and editing [29].

To validate this, we conducted an inversion and resampling experiment on TRELLIS-generated assets using three prompt conditions: the true prompt, a two-word approximate variant, and an empty prompt. Reconstruction accuracy

was evaluated using the L_1 score on the coarse voxel structure produced by the geometric stage $\mathcal{G}_S$. As shown in Tab. 2, approximate prompts led to poor reconstructions in TRELLIS, whereas in comparison trials with Stable Diffusion [34] and FLUX.1 [17], reconstructions with approximate prompts remained comparable to those using original prompts.

We further analyze this instability by computing the norm of the predicted velocity during inversion. In T2I models, the sequence of score norms across noise levels can detect OOD data samples [28]. For the flow-based formulation of TRELLIS, we evaluate the equivalent sequence of velocity norms $\|\mathbf{v}_t\|$ across timesteps [18]. Our results show that approximate prompts result in substantially higher velocity norms, signaling an attempt by the model to escape low-density regions.

This analysis highlights both the difficulty of 3D-based inversion and the lack of previous purely text-driven editing methods. Indeed, as we show, because they become OOD prompts, using approximate prompts as advocated in image-based approaches [29], is not suited for text-to-3D models, even state-of-the-art ones like TRELLIS.

Instead, we find that the *empty prompt* instead promotes substantially smaller velocity norms and more stable trajectories (Fig. 4 and Table 2). In classifier-free guidance, the w -difference term between conditional and unconditional velocities can act as an unstable, high-variance gradient when the conditioning is highly out-of-distribution (*e.g.*, from an approximate prompt). Using an empty prompt eliminates this OOD guidance, keeping the trajectory stable. This suggests that combining the null prompt with a shape is a more reliable strategy for embedding such pairs into the model’s latent space, and also enables stable text-based editing when combined with inversion.

4.3 Geometric Expressivity and Text-Editing

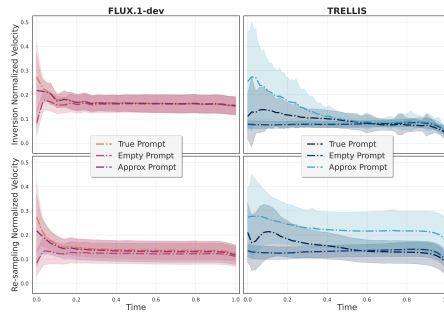


Fig. 4: The velocity norm of FLUX.1 [17] remains stable across different prompt types, whereas TRELLIS [43] exhibits large variations.

Table 2: Effect of prompt types on inversion quality across different models.

	True Prompt	Approximate Prompt	Empty Prompt
Stable Diffusion v1.4 [34]			
PSNR $\uparrow$	15.00 $\pm$ 2.43	14.61 $\pm$ 3.45	15.66 $\pm$ 2.57
LPIPS $\downarrow$	0.56 $\pm$ 0.07	0.59 $\pm$ 0.08	0.57 $\pm$ 0.07
FLUX.1 dev [17]			
PSNR $\uparrow$	10.32 $\pm$ 2.55	10.62 $\pm$ 6.12	15.12 $\pm$ 3.33
LPIPS $\downarrow$	0.57 $\pm$ 0.169	0.58 $\pm$ 0.19	0.48 $\pm$ 0.13
TRELLIS [43]			
L1 $\downarrow$	17.75 $\pm$ 34.59	76.55 $\pm$ 73.35	5.40 $\pm$ 14.15
LN3Diff [19]			
L1 $\downarrow$	1.75 $\pm$ 0.98	1.82 $\pm$ 1.16	0.05 $\pm$ 0.07
GaussianAnything [20]			
L1 $\downarrow$	0.32 $\pm$ 0.45	0.99 $\pm$ 0.47	0.19 $\pm$ 0.06

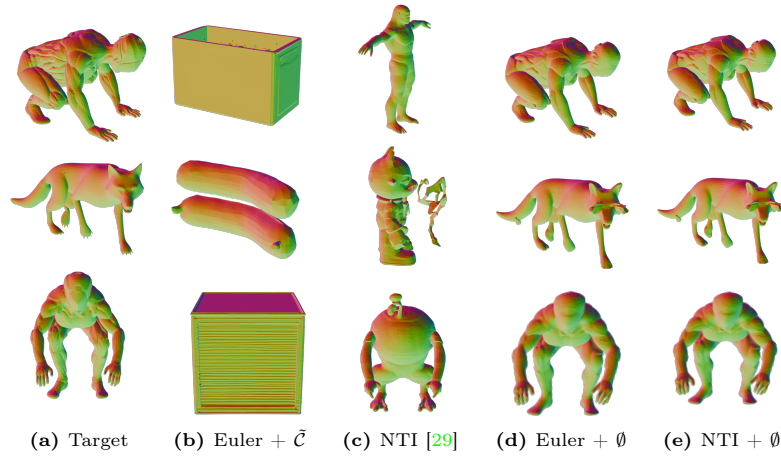


Fig. 5: Comparison of shape reconstruction across inversion methods. (a) Target shape. (b) Euler inversion with approximate text prompt. (c) NTI inversion with approximate text prompt. (d) Euler inversion with empty prompt $\emptyset$. (e) NTI inversion with empty prompt $\emptyset$. When using approximate text prompts (b-c), both methods fail to accurately reconstruct the target shape. In contrast, inverting with an empty prompt (d-e) enables both methods to achieve high-fidelity reconstruction.

Building on the finding that null prompts coupled with shape embeddings mitigate OOD drift, we utilize this stability to probe the model’s inherent geometric expressivity. As illustrated in Fig. 5 and in the experiment section 5, using an approximate prompt $\tilde{C}$ often leads to implausible geometries or artifacts, likely due to the latent trajectory drifting into sparse density regions during inversion. This capability demonstrates that unconditional embeddings can guide the model to recreate complex shapes where finding the correct text-conditioning would be very challenging. We interpret the phenomenon as a clear “expressivity mismatch”: the model’s geometric capacity far exceeds its language-constrained expressivity, proving that the latent space contains a wealth of geometric detail accessible only through unconditional guidance.

Text-guided editing of arbitrary shapes. Our text-driven editing pipeline leverages the unconditional distribution (see Fig. 3 for an overview). Because we invert arbitrary shapes, our approach enables text-based retargeting. Namely, given a shape $\mathcal{X}$, the objective is to re-target the source pose to diverse target characters (see Fig. 6). For example, if the source shape is a demon figure dancing, we can use edit prompts such as $\mathcal{P}_{edit}$ “princess dancing” or “cowboy dancing”. We first invert $\mathcal{X}$ into the TRELIS latent space using a null prompt ($\mathcal{P}_{src} = \emptyset$). To perform an edit, we resample from the inverted noise latent using the edit prompt ($\mathcal{P}_{edit}$). We optionally apply null-text optimization [29] to optimize the unconditional embeddings along the sampling trajectory, which allows the edit to maintain further structural integrity to the original shape.



Fig. 6: Open-vocabulary edits (right) on non-rigid 3D shapes (left) inverted by our method within the unconditional latent space of the TRELLIS 3D generative model. Our inversion method reliably reconstructs arbitrary shapes and enables semantic edits.

5 Experiments

5.1 Experimental setup

Implementation For inversion, we use classifier-free guidance (CFG) scale of 5 and 50 sampling timesteps. We perform inversion exclusively on the first-stage flow model $\mathcal{G}_S$, which dictates the global geometric structure, while keeping the second-stage model $\mathcal{G}_L$ fixed and seeded to isolate the source of variation. When employing Null-Text Inversion (NTI) [29], we use the Adam optimizer [15] with a learning rate of 1×10^{-4} and 10 inner optimization steps per sampling step, conducted on a single L40S GPU. We measure reconstruction quality using an L1 metric on the decoded latents, defined as $L1 = |\hat{\mathbf{x}}_0 - \mathcal{D}(\mathcal{E}(\mathbf{x}_0))|$, where $\mathcal{E}$ and $\mathcal{D}$ represent the VAE encoder and decoder. This ensures we account for the inherent VAE reconstruction bias. To capture finer surface details and perceptual fidelity, we also compute LPIPS [48] scores between rendered normals of the original and reconstructed meshes across 10 distinct viewpoints.

Datasets We evaluate our method using two datasets: 200 non-rigid humanoid and animal characters from the DT4D dataset [23] to assess open-vocabulary semantic edits on complex, out-of-distribution (OOD) shapes, and a subset of 80 TRELLIS-generated shapes. The latter is used strictly for comparisons to other methods, as existing methods like VoxHammer [21] suffer from latent value explosion on DT4D (see Supp. Sec. D for precise dataset preparation). The detailed creation of the TRELLIS-generated shapes, as well as the setups of Sections 4.1 and 4.2 are detailed in the supplementary material.



Fig. 7: Comparison of our editing with native 3D space baselines: the state-of-the-art VoxHammer, and the TRELIS second stage edit. VoxHammer fails to create plausible shapes due to the inversion failure, while TRELIS proposes only texture changes. Ours is the only method to provide meaningful retargeting based on the editing prompts.

5.2 Evaluation of geometric expressivity

To evaluate the geometric expressivity of TRELIS, we compare four configurations: Euler inversion and NTI, each paired with either an approximate prompt $\tilde{C}$ (e.g., “A horse galloping”) or an empty prompt $\emptyset$. Results in Supp. Mat. (Table 1) demonstrate that using an approximate prompt can be detrimental to reconstruction quality; while NTI with $\tilde{C}$ improves reconstruction over Euler with $\tilde{C}$, it remains significantly worse than using an empty prompt. Notably, when using $\emptyset$, both Euler and NTI achieve similar, high-accuracy results.

As illustrated in Fig. 5, using an approximate prompt $\tilde{C}$ often leads to implausible geometries or artifacts, likely due to the latent trajectory drifting into sparse density regions during inversion. In contrast, Euler inversion with a null prompt consistently achieves the highest reconstruction accuracy, successfully capturing the complex, non-rigid topologies inherent in DT4D. This capability demonstrates that unconditional embeddings can guide the model to recreate complex shapes where finding in-distribution text-conditioning would be very challenging.

5.3 Text-Guided Editing Evaluation

We now evaluate editing capabilities by creating edit prompts $\mathcal{P}_{edit}$ that modify appearance while preserving the original pose and structure.

Baselines. We compare our approach against the native editing procedure of TRELIS [43], the text-based pipeline of VoxHammer [21], and two image-prior baselines: Vox-E [36] and Instant3DiT [1] (visual comparisons in Supp. Fig. 7. We



Fig. 8: Multiview consistency of 3D edits. Because our method operates natively in the 3D space, the provided edits are naturally view-consistent.

exclude full-scene editing methods to ensure a fair, direct quantitative comparison with our mesh-based object editing. Additionally, TRELIS’ native editing retains coarse structures by re-running only its second-stage model ($\mathcal{G}_L$); however, it primarily modifies texture with negligible geometric updates (Fig. 7). Furthermore, VoxHammer fails on 67% of DT4D dataset samples due to inversion drift causing latent value explosion (rendering shapes undecodable by the VAE). Even when successful, VoxHammer requires manual 3D masking, frequently produces implausible geometry, and struggles to modify subjects while preserving their original pose.

Quantitative and Visual Results. To evaluate editing quality, we sample 30 rendered views per edited sample. Table 3 reports alignment with the target prompt $\mathcal{P}_{edit}$ using SigLIP [39], CLIP scores, and LLaVA-1.5-7B [25] for VLM-based evaluation (1–5 scale). Starting from inverted noise (Euler + $\emptyset$ and NTI + $\emptyset$), our method consistently outperforms all baselines. Beyond visual quality, our approach offers a substantial computational advantage. Both Instant3DiT [1] and VoxHammer [21] require manual mask creation, with VoxHammer additionally rendering over 150 views to extract multi-view DINOv2 [30] features. Meanwhile, Vox-E [36] relies on notoriously slow per-scene optimization, taking over an hour per edit. In contrast, our pipeline requires only input mesh voxelization and three generative sampling passes (averaging 9s per edit). This drastically improves computational efficiency while maintaining higher structural integrity.

5.4 Geometric Consistency

Despite enabling global geometric modifications, our edits remain strictly 3D-consistent. We provide multiview renders of our results in Fig. 8. These renders confirm that our method maintains structural coherence and rendering stability across all viewing angles. Further videos are provided in the supplementary material.

6 Limitations and future work

Limitations. Although our method produces high-quality edits, it is fundamentally bounded by the generative model’s underlying distribution. Unconditional inversion successfully recovers the structure of highly out-of-distribution



Fig. 9: Bottom-left: input shape in a dancing pose on the floor. Bottom-right: two edits. The edits showcase failure modes, likely because the pose is geometrically out of distribution.

Table 3: Quantitative evaluation of the edited assets. $\dagger$ indicates methods requiring image-based priors.

Method	SigLIP $\uparrow$	CLIP $\uparrow$	VLM $\uparrow$
TRELLIS.1 [43]	0.0797	0.2489	2.65
VoxHammer [21]	0.0240	0.2207	1.33
Vox-E † [36]	0.0250	0.2430	1.84
Instant3DiT † [1]	0.0405	0.2434	1.99
Ours	0.1469	0.2829	3.98

shapes, but the subsequent editing trajectories can become unstable under simple prompts. In these extreme cases, applying edit prompts with the retrieved noise latent and unconditional embeddings yields implausible geometries (Fig. 9). Because our optimization does not explicitly enforce geometric plausibility, large semantic shifts on rare poses may produce unrealistic results.

Future work. First, future research can focus on improving editing robustness, notably by automatically detecting out-of-distribution and low-diversity states. Second, inversion could be applied to new downstream tasks; since it captures rich geometric cues, adapting image morphing techniques [45] to 3D shapes would benefit shape analysis applications.

7 Conclusion

We present the first systematic investigation of inversion techniques for native 3D generative models, revealing fundamental differences from established image-domain methods. Our key finding is that weak text-shape alignment in 3D generative models causes standard prompt-guided inversion to drift into sparse regions of the learned distribution, resulting in poor reconstructions and implausible geometries. To address this challenge, we propose a simple yet effective solution: leveraging the unconditional generative prior during inversion. By performing Euler inversion with an empty prompt and optimizing only the unconditional embeddings via null-text optimization, we achieve high-fidelity reconstructions of complex, out-of-distribution shapes. Remarkably, this unconditional inversion strategy not only improves reconstruction quality but also enables meaningful text-driven editing while preserving the source shape’s global structure and pose. Finally, even with the rapid growth of 3D data repositories, text-to-3D generation is likely to remain data-constrained, meaning user-provided meshes may lack corresponding generative prompts. Our approach sheds light on the possibility of inversion and editing in such data-constrained regimes.

Acknowledgements

Parts of this work were supported by the ERC Consolidator Grant “VEGA” (No. 101087347), as well as gifts from Ansys Inc., and Adobe Research. This project was provided with computing AI and storage resources by GENCI at IDRIS thanks to the grant 2026-AD010613760R3 on the supercomputer Jean Zay’s H100 partition.

References

1. Barda, A., Gadelha, M., Kim, V.G., Aigerman, N., Bermano, A.H., Groueix, T.: Instant3dit: Multiview inpainting for fast editing of 3d objects (2024), <https://arxiv.org/abs/2412.00518>
2. Chen, Y., Pan, Y., Li, Y., Yao, T., Mei, T.: Control3d: Towards controllable text-to-3d generation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1148–1156 (2023)
3. Chihaoui, H., Lemkhenter, A., Favaro, P.: Blind image restoration via fast diffusion inversion. *Advances in Neural Information Processing Systems* **37**, 34513–34532 (2024)
4. Dinh, N.A., Lang, I., Kim, H., Stein, O., Hanocka, R.: Geometry in style: 3d stylization via surface normal deformation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28456–28467 (2025)
5. Erkoç, Z., Gümeli, C., Wang, C., Niekner, M., Dai, A., Wonka, P., Lee, H.Y., Zhuang, P.: Predictor3d: Fast and precise 3d shape editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 640–649 (2025)
6. Gao, W., Wang, D., Fan, Y., Bozic, A., Stuyck, T., Li, Z., Dong, Z., Ranjan, R., Sarafianos, N.: 3d mesh editing using masked lrms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7154–7165 (2025)
7. Gemini Team Google: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (July 2025)
8. Google: Gemini 3 pro: Our most advanced reasoning model (November 2025), <https://blog.google/products-and-platforms/products/gemini/gemini-3/>, accessed: 2026-03-03
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
10. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
11. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(6), 4409–4437 (Jun 2025). <https://doi.org/10.1109/tpami.2025.3541625>, <http://dx.doi.org/10.1109/TPAMI.2025.3541625>
12. Jiao, G., Huang, B., Wang, K.C., Liao, R.: Uniedit-flow: Unleashing inversion and editing in the era of flow models. arXiv preprint arXiv:2504.13109 (2025)
13. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions (2023), <https://arxiv.org/abs/2305.02463>

14. Kim, H., Lang, I., Aigerman, N., Groueix, T., Kim, V.G., Hanocka, R.: Meshup: Multi-target mesh deformation via blended score distillation. In: 2025 International Conference on 3D Vision (3DV). pp. 222–239. IEEE (2025)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017), <https://arxiv.org/abs/1412.6980>
16. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
17. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), <https://arxiv.org/abs/2506.15742>
18. Lai, C.H., Song, Y., Kim, D., Mitsufuji, Y., Ermon, S.: The principles of diffusion models (2025), <https://arxiv.org/abs/2510.21890>
19. Lan, Y., Hong, F., Zhou, S., Yang, S., Meng, X., Chen, Y., Lyu, Z., Dai, B., Pan, X., Loy, C.C.: Ln3diff++: Scalable latent neural fields diffusion for speedy 3d generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **48**(3), 3002–3017 (Mar 2026). <https://doi.org/10.1109/tpami.2025.3633073>, <http://dx.doi.org/10.1109/TPAMI.2025.3633073>
20. Lan, Y., Zhou, S., Lyu, Z., Hong, F., Yang, S., Dai, B., Pan, X., Loy, C.C.: Gaussiananything: Interactive point cloud flow matching for 3d object generation (2025), <https://arxiv.org/abs/2411.08033>
21. Li, L., Huang, Z., Feng, H., Zhuang, G., Chen, R., Guo, C., Sheng, L.: Voxhammer: Training-free precise and coherent 3d editing in native 3d space. *arXiv preprint arXiv:2508.19247* (2025)
22. Li, P., Ma, S., Chen, J., Liu, Y., Zhang, C., Xue, W., Luo, W., Sheffer, A., Wang, W., Guo, Y.: Cmd: Controllable multiview diffusion for 3d editing and progressive generation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025)
23. Li, Y., Takehara, H., Taketomi, T., Zheng, B., Niessner, M.: 4dcomplete: Non-rigid motion estimation beyond the observable surface. *IEEE International Conference on Computer Vision (ICCV)* (2021)
24. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023), <https://arxiv.org/abs/2210.02747>
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), <https://arxiv.org/abs/2304.08485>
26. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022)
27. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9970–9980 (2024)
28. Mahmood, A., Oliva, J., Styner, M.: Multiscale score matching for out-of-distribution detection (2021), <https://arxiv.org/abs/2010.13132>
29. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6038–6047 (2023)

30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), <https://arxiv.org/abs/2304.07193>
31. Parelli, M., Oechsle, M., Niemeyer, M., Tombari, F., Geiger, A.: 3d-latte: Latent space 3d editing from textual instructions (2025), <https://arxiv.org/abs/2509.00269>
32. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023), <https://openreview.net/forum?id=FjNys5c7VyY>
33. Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies (2024), <https://arxiv.org/abs/2312.03806>
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), <https://arxiv.org/abs/2112.10752>
35. Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., Chu, W.S.: Semantic image inversion and editing using rectified stochastic differential equations (2024), <https://arxiv.org/abs/2410.10792>
36. Sella, E., Fiebelman, G., Hedman, P., Averbuch-Elor, H.: Vox-e: Text-guided voxel editing of 3d objects (2023), <https://arxiv.org/abs/2303.12048>
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=StIgiarCHLP>
38. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32** (2019)
39. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harnsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025), <https://arxiv.org/abs/2502.14786>
40. Wallace, B., Gokul, A., Naik, N.: Edict: Exact diffusion inversion via coupled transformations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22532–22541 (2023)
41. Wang, P., Xu, D., Fan, Z., Wang, D., Mohan, S., Iandola, F., Ranjan, R., Li, Y., Liu, Q., Wang, Z., et al.: Taming mode collapse in score distillation for text-to-3d generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9037–9047 (2024)
42. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., Yang, J.: Native and compact structured latents for 3d generation. *Tech report* (2025)
43. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. *CVPR* (2024)
44. Xu, P., Jiang, B., Hu, X., Luo, D., He, Q., Zhang, J., Wang, C., Wu, Y., Ling, C., Wang, B.: Unveil inversion and invariance in flow transformer for versatile image editing. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 28479–28489 (2025)

45. Yang, Z., Yu, Z., Xu, Z., Singh, J., Zhang, J., Campbell, D., Tu, P., Hartley, R.: IMPUS: Image morphing with perceptually-uniform sampling using diffusion models. In: The Twelfth International Conference on Learning Representations (2024), <https://openreview.net/forum?id=gG38EBE2S8>
46. Ye, J., Xie, S., Zhao, R., Wang, Z., Yan, H., Zu, W., Ma, L., Zhu, J.: Nano3d: A training-free approach for efficient 3d editing without masks (2025), <https://arxiv.org/abs/2510.15019>
47. Zhang, B., Cheng, Y., Yang, J., Wang, C., Zhao, F., Tang, Y., Chen, D., Guo, B.: Gaussiancube: A structured and explicit radiance representation for 3d generative modeling (2024), <https://arxiv.org/abs/2403.19655>
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
49. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10146–10156 (2023)