ENTRO-BRIDGE: ENTROPY-STRUCTURED REFINEMENT FOR FEED-FORWARD 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in single-view 3D reconstruction have been enabled by feed-forward and generative models, allowing efficient learning-based inference from incomplete visual observations. However, these models often overlook the spatially heterogeneous reliability of reconstructed geometry: feed-forward predictions may contain reliable observations, uncertain predictions, and unsupported regions, while generative completion may overwrite reliable structures during unseen-region recovery. We formulate this challenge as a reliability-aware stochastic repair problem, where reliable geometry should be preserved, uncertain predictions should be corrected, and unsupported regions should be completed. To solve this problem, we propose Entro-Bridge, an entropy-structured paired Schrödinger Bridge framework that enables selective geometric repair through uncertainty-controlled stochastic transport. Specifically, we introduce an Entropy-structured Reliability-aware Source Boundary (ERSB), which transforms feed-forward predictions into a source distribution with spatially structured entropy: reliable observations are assigned low uncertainty, uncertain predictions allow greater flexibility for correction, and unsupported regions retain greater freedom for completion. During reverse transport, the Reliability-Guided reverse Repair Dynamics (RGRD) carries this reliability structure into the reverse repair process, guiding the correction trajectory toward uncertain regions while preserving reliable geometry. On the ABO and 3D-FRONT datasets, Entro-Bridge consistently outperforms state-of-the-art methods, reducing Chamfer Distance by approximately 13.0% and 10.3% relative to SOTA methods, with F-scores of 92.0% and 81.3%, respectively. Our project website is available at https://entrobridge.github.io/Entro-Bridge/.

1 INTRODUCTION

Single-view 3D reconstruction aims to recover complete 3D geometry from a single RGB image, requiring models to infer structured 3D representations from highly incomplete visual information. This problem lies at the intersection of representation learning and generative modeling, where recent advances in learned 3D representations and generative models have substantially improved reconstruction quality and geometric plausibility (Mildenhall et al., 2021; Kerbl et al., 2023; Lan et al., 2025; Xiang et al., 2025). Compared with multi-view reconstruction, single-view inference greatly reduces acquisition requirements, making it particularly attractive when observations are sparse or difficult to obtain.

Existing feed-forward reconstruction methods such as DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), and VGGT (Wang et al., 2025) enable efficient geometry prediction. However, a single view leaves back-facing and occluded regions unsupported, while visible predictions may mix reliable geometry with errors caused by weak texture, reflections, occlusion boundaries, and depth ambiguity. Generative completion methods, including 3D-EPN (Dai et al., 2017), cGAN (Wu et al., 2020), AutoSDF (Mittal et al., 2022), and ShapeFormer (Yan et al., 2022), can infer missing structures, yet typically treat partial geometry as a condition without explicitly distinguishing reliable evidence from uncertain predictions. As a result, they may overwrite reliable observations or introduce inconsistencies near observed–unobserved transitions (Fig. 1). We therefore formulatefeed-forward

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.

reconstruction refinement as a reliability-aware stochastic repair problem: reliable geometry should be preserved, uncertain predictions should be corrected, and unsupported regions should be completed within a unified process.

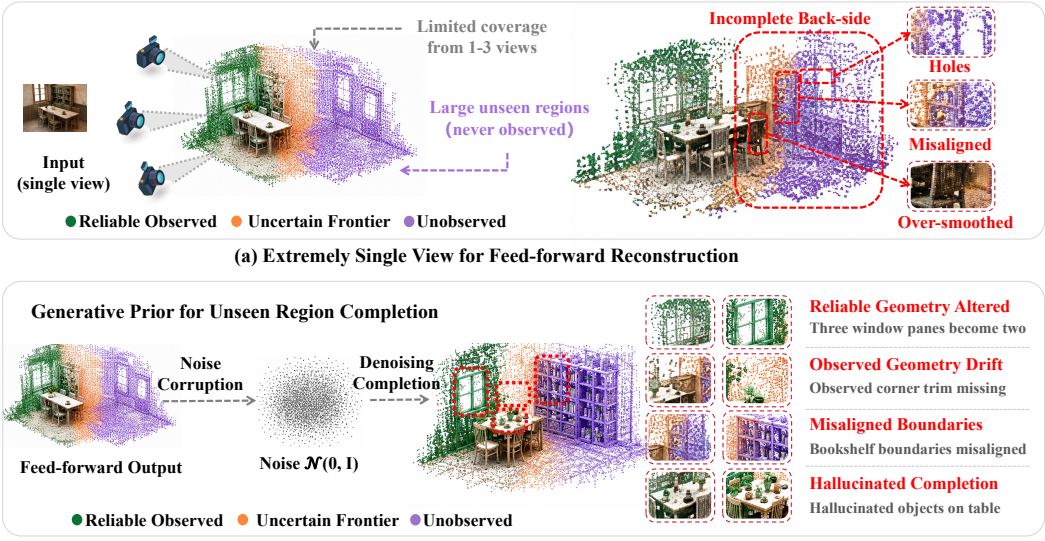

Figure 1: **Motivation under the single-view setting.** Feed-forward reconstruction is spatially unreliable, while generative completion may alter reliable geometry.

Schrödinger Bridges (De Bortoli et al., 2021) provide entropy-regularized transport between endpoint distributions, while I^2SB (Liu et al., 2023) shows that degraded observations can serve as informative restoration sources. These properties make paired stochastic transport suitable for reliability-aware repair, provided that the feed-forward prediction is not treated as a uniformly degraded endpoint. We therefore use incomplete feed-forward reconstruction as an informative, reliability-structured bridge source rather than Gaussian noise, allowing different spatial regions to retain different degrees of correction freedom.

We propose Entro-Bridge, a reliability-aware paired Schr”odinger Bridge framework for single-view 3D reconstruction. Its Entropy-structured Reliability-aware Source Boundary (ERSB) transforms feed-forward predictions into a source distribution whose spatial entropy reflects observation support and confidence: reliable regions remain constrained, uncertain regions allow greater correction, and unsupported regions permit completion. Reliability-Guided Reverse Repair Dynamics (RGRD) propagates this structure through reverse transport, focusing correction on uncertain regions while preserving reliable geometry. Evaluations on ABO and 3D-FRONT further demonstrate the effectiveness of Entro-Bridge, reducing Chamfer Distance by approximately 13.0% and 10.3% relative to SOTA methods on ABO and 3D-FRONT, with F-scores of 92.0% and 81.3%, respectively.

Our method makes three main technical contributions:

- We formulate feed-forward refinement as reliability-aware stochastic repair, distinguishing reliable, uncertain, and unsupported regions for preservation, correction, and completion. ERSB constructs the corresponding spatially varying source uncertainty from observation support and confidence.
- We develop RGRD, which uses region-aware supervision and reliability-guided local propagation to preserve reliable geometry while repairing uncertain and unsupported regions.
- Under supervised endpoint coupling, we instantiate this formulation with a paired conditional bridge, reducing the path-space KL objective to tractable conditional Gaussian bridges for efficient residual learning.

2 Related Work

2.1 Feed-forward 3D Reconstruction

Feed-forward 3D reconstruction avoids per-scene optimization (Schonberger & Frahm, 2016; Yao et al., 2018; 2019; Yu & Gao, 2020). DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) predict point maps or geometric correspondences, while Gaussian-based methods such as TranSplat (Zhang et al., 2025), HiSplat (Tang et al., 2025), Generative Densification (Nam et al., 2025), NoPoSplat (Ye et al., 2025), UniSplat (Shi et al., 2026), and Fuse3D (Jin et al., 2025) improve explicit reconstruction, alignment, and sparse-view fusion. Related studies have also explored hierarchical features and state-space modeling for local global aggregation (Zhao et al., 2026a;b). Yet single-view predictions still exhibit structural drift, over-smoothing, and missing geometry. DIFIX3D (Wu et al., 2025a), One-shot Refiner (Dong et al., 2026), and S2D (Ji et al., 2026) add generative refinement. ReconViaGen (Chang et al., 2026) integrates VGGT reconstruction priors into TRELLIS as multi-view conditions for 3D generation. In contrast, Entro-Bridge treats the incomplete feed-forward geometry itself as a reliability-structured bridge source and learns source-to-complete transport directly in 3D latent space.

2.2 3D Shape Completion and Generative Diffusion Models

3D shape completion aims to recover the complete geometry from partial observations, corrupted scans, or incomplete structural cues. Early methods (Dai et al., 2017; Han et al., 2017) formulate completion as a deterministic mapping from partial inputs to complete shapes, yet often struggle with ambiguous missing regions and produce over-smoothed results. Generative methods, including cGAN (Wu et al., 2020), AutoSDF (Mittal et al., 2022), and ShapeFormer (Yan et al., 2022), improve shape-distribution modeling, yet may generate plausible structures that are not strictly consistent with the observed input.

More recently, diffusion and latent generative models have become important paradigms for 3D generation and completion. TRELLIS (Xiang et al., 2025), GaussianAnything (Lan et al., 2025), and LGM (Tang et al., 2024) improve high-quality 3D generation in compact latent spaces, while Real3D (Jiang et al., 2025), VoxHammer (Li et al., 2026), SceneGen (Meng et al., 2026), and MIDI (Huang et al., 2025) extend generative 3D modeling toward real-data alignment, controllable editing, and scene-level generation. Related 3D Gaussian-based frameworks further explore text-guided object editing and compositional 3D generation (Xue et al., 2026; Tian et al., 2026). Amodal3R (Wu et al., 2025b) augments TRELLIS with visibility- and occlusion-aware attention, while DiffusionGS (Cai et al., 2025) directly predicts 3D Gaussians during denoising. BridgeShape (Kong et al., 2026) applies a Schrödinger Bridge to transport partial shape latents toward complete shapes, whereas Entro-Bridge explicitly models the spatially varying reliability of feed-forward geometry through ERSB and reliability-guided repair.

3 Our Approach

Entro-Bridge formulates feed-forward reconstruction refinement as reliability-aware stochastic repair: reliable, uncertain, and unsupported regions require preservation, correction, and completion, respectively. ERSB constructs a reliability-structured source boundary, while RGRD carries this structure into reverse repair on a paired latent bridge. The overall pipeline is illustrated in Fig. 2.

3.1 Reliability-Aware Repair Formulation and Paired Latent Bridge

Given an input image $\mathbf{I}$, a frozen feed-forward reconstructor predicts the incomplete geometry $\mathcal{P}_{\text{obs}}$ and its associated confidence. We voxelize $\mathcal{P}_{\text{obs}}$ and the corresponding complete ground-truth geometry $\mathcal{P}_{\text{gt}}$ and encode both with the frozen TRELLIS sparse-structure encoder (Xiang et al., 2025):

$$\mathbf{q}_s = E(V(\mathcal{P}_{\text{obs}})), \qquad \mathbf{q}_g = E(V(\mathcal{P}_{\text{gt}})), \tag{1}$$

where $V(\cdot)$ denotes voxelization or sparse-structure conversion, $E(\cdot)$ is the frozen TRELLIS sparse-structure encoder, and $\mathbf{q}_s, \mathbf{q}_g \in \mathbb{R}^{c_s \times 16 \times 16 \times 16}$ denote the spatially aligned source and target latent tensors, respectively, with c_s denoting the latent channel dimension.

We represent source reliability on the aligned latent grid using a binary observation-support map $\mathbf{M} = \{M_i\}_{i=1}^{16^3}$ with $M_i \in \{0, 1\}$ and a confidence field $\mathbf{C} = \{C_i\}_{i=1}^{16^3}$ with $C_i \in [0, 1]$. Here, $M_i = 1$ indicates that latent cell i is supported by the feed-forward reconstruction, while C_i measures the reliability of that support; unsupported cells are assigned $C_i = 0$. Both $\mathbf{M}$ and $\mathbf{C}$ are single-channel fields aligned with the 16^3 latent grid and are broadcast over the latent-channel dimension in element-wise operations.

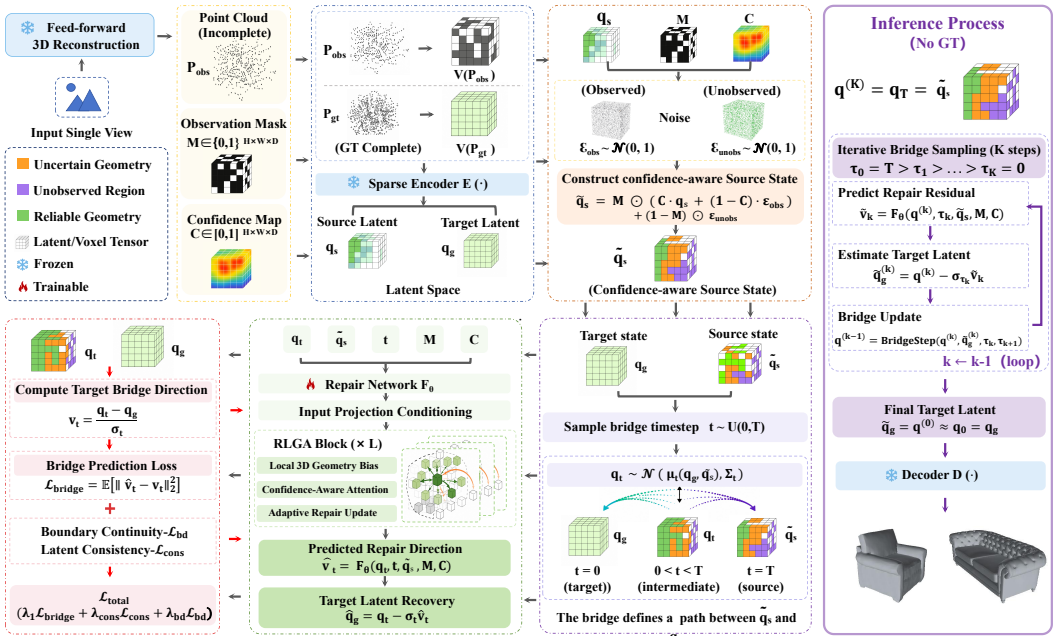

Figure 2: **Overview of Entro-Bridge.** Observation support and confidence define the ERSB, preserving reliable evidence while assigning greater corrective freedom to uncertain regions and generative freedom to unsupported regions. A paired latent bridge provides the transport backbone, while region-aware supervision and RLGA carry the same reliability structure into the learned reverse repair process.

Reliability-aware stochastic repair formulation. Together, $\mathbf{M}$ and $\mathbf{C}$ characterize the spatially heterogeneous reliability of the feed-forward source: high-confidence supported cells should be preserved, low-confidence observed cells corrected, and unsupported cells completed. We therefore map a reliability-structured source measure $p_s^{\mathbf{M},\mathbf{C}}$ toward the complete-geometry distribution p_g, with ERSB defining the source measure and RGRD realizing the corresponding selective repair during reverse reconstruction.

Given the reliability-structured source measure defined above, we use a Schrödinger Bridge (SB) as the stochastic transport backbone. The bridge is the path-space relative-entropy projection

$$\mathbb{P}^\star = \arg\min_{\mathbb{P}} D_{\mathrm{KL}}(\mathbb{P} \| \mathbb{R}_{\mathrm{ref}}), \quad \mathbb{P}_0 = p_g,\ \mathbb{P}_T = p_s^{\mathbf{M},\mathbf{C}}, \tag{2}$$

where $\mathbb{P}$ denotes a candidate probability measure over latent trajectories $\mathbf{q}_{0:T} = \{\mathbf{q}_t\}_{t=0}^{T}$, with $\mathbf{q}_t$ denoting the latent bridge state at timestep t and T denoting the terminal bridge time, i.e., the total number of reference-diffusion timesteps. $\mathbb{P}^\star$ is the resulting optimal bridge path measure, and $\mathbb{R}_{\mathrm{ref}}$ denotes the path measure of the reference diffusion process. $\mathbb{P}_0$ and $\mathbb{P}_T$ denote the endpoint marginals of $\mathbb{P}$, p_g is the distribution of complete-geometry latents, and $p_s^{\mathbf{M},\mathbf{C}}$ is the reliability-structured source measure associated with the support and confidence fields $\mathbf{M}$ and $\mathbf{C}$. The path-space Kullback–Leibler divergence D_{KL} provides the classical entropy regularization of the Schrödinger Bridge.

Paired conditional reduction. Unlike the generic marginal-only problem, supervised training provides paired endpoints $(\mathbf{q}_0, \mathbf{q}_T) = (\mathbf{q}_g, \tilde{\mathbf{q}}_s)$, where $\tilde{\mathbf{q}}_s$ denotes the stochastic reliability-aware

source endpoint obtained from $\mathbf{q}_s$ using $\mathbf{M}$ and $\mathbf{C}$. These paired endpoints fix an empirical endpoint coupling π_{pair}, i.e., the empirical joint distribution over paired complete and source endpoints. By the KL chain rule, the path-space objective decomposes into an endpoint-coupling term, constant under π_{pair}, plus the expected conditional path KL. Hence the paired entropy projection reduces to the reference bridge conditioned on each endpoint pair; Appendix A.1 gives the derivation.

For the drift-free Gaussian reference diffusion used here, this conditional bridge has the tractable form (Liu et al., 2023; Kong et al., 2026)

$$\mathbf{q}_t = \mu_{0,t}\mathbf{q}_g + \mu_{1,t}\tilde{\mathbf{q}}_s + \sigma_{\text{br},t}\boldsymbol{\epsilon}. \tag{3}$$

Here, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{\text{lat}})$ is standard Gaussian noise, $\mu_{0,t}$ and $\mu_{1,t}$ are the time-dependent interpolation coefficients of the complete and source endpoints, respectively, $\sigma_{\text{br},t}$ denotes the conditional bridge standard deviation, and $\mathbf{I}_{\text{lat}}$ denotes the identity covariance matrix with dimensionality matching the latent state. The reference process is indexed from complete geometry ($t = 0$) to the source endpoint ($t = T$), whereas reconstruction follows the learned reverse conditional bridge. The closed-form interpolation coefficients, covariance, and variance schedule are given in Appendix A.

Reverse-bridge learning. For a sampled bridge state, we define the normalized target-side residual as

$$\mathbf{v}_t = \frac{\mathbf{q}_t - \mathbf{q}_g}{\sigma_t}, \qquad \hat{\mathbf{v}}_t = F_\theta(\mathbf{q}_t, t, \tilde{\mathbf{q}}_s, \mathbf{M}, \mathbf{C}), \tag{4}$$

where $\mathbf{v}_t$ denotes the normalized residual from the complete endpoint $\mathbf{q}_g$ to the current bridge state $\mathbf{q}_t$, and $\hat{\mathbf{v}}_t$ is its prediction by the repair network F_θ, parameterized by θ. Here, σ_t is the accumulated reference standard deviation from the complete endpoint to timestep t.

The bridge regression objective and the corresponding target-endpoint estimate $\hat{\mathbf{q}}_g$ are

$$\mathcal{L}_{\text{bridge}} = \mathbb{E}\left[\left\| F_\theta(\mathbf{q}_t, t, \tilde{\mathbf{q}}_s, \mathbf{M}, \mathbf{C}) - \frac{\mathbf{q}_t - \mathbf{q}_g}{\sigma_t} \right\|_2^2\right], \qquad \hat{\mathbf{q}}_g = \mathbf{q}_t - \sigma_t F_\theta(\mathbf{q}_t, t, \tilde{\mathbf{q}}_s, \mathbf{M}, \mathbf{C}), \tag{5}$$

where the expectation is taken over paired training endpoints, the sampled bridge timestep t, and the Gaussian perturbations used to construct the stochastic source and intermediate bridge states. $\hat{\mathbf{q}}_g$ denotes the estimated complete target endpoint recovered from the current bridge state.

For the bridge regression term in isolation, the residual loss equals $\sigma_t^{-2}\|\hat{\mathbf{q}}_g - \mathbf{q}_g\|_2^2$, and its population minimizer yields the conditional-mean endpoint estimator under the sampled paired bridge. Thus, equation 5 provides a bridge-consistent reverse endpoint estimator rather than solving the generic marginal SB directly; the region-aware regularizers introduced below intentionally shape this estimator toward task-specific preservation, correction, and completion. Appendix A.1 gives the full argument.

At inference, we initialize from $\mathbf{q}_T = \tilde{\mathbf{q}}_s$ and use K deterministic DDIM-style updates, where K denotes the number of reverse-bridge sampling steps. These updates are constructed from the learned endpoint estimator and the conditional-bridge coefficients while reusing the recovered noise direction; the exact sampler is given in Appendix A.

Classical SB regularizes stochastic transport through path-space relative entropy, yet does not specify how uncertainty should be spatially allocated at the source endpoint. We therefore introduce confidence-induced endpoint entropy to structure local uncertainty before transport.

3.2 Entropy-Structured Reliability-Aware Source Boundary

The feed-forward source is not a clean partial shape: its reliability varies spatially. We therefore seek a stochastic source measure satisfying two reliability-consistent properties. First, source retention should increase with confidence, so highly reliable observations remain close to the feed-forward prediction. Second, source-endpoint uncertainty should decrease with confidence, so uncertain observations retain greater corrective freedom. Unsupported regions, for which no source-specific evidence is available, should instead revert to the Gaussian prior at the source boundary.

To jointly instantiate source retention and reliability-dependent uncertainty, we parameterize

$$\tilde{\mathbf{q}}_s = \mathbf{M} \odot \big(\mathbf{C} \odot \mathbf{q}_s + (1 - \mathbf{C}) \odot \boldsymbol{\epsilon}_{\text{obs}}\big) + (1 - \mathbf{M}) \odot \boldsymbol{\epsilon}_{\text{unobs}}. \tag{6}$$

Here, $\boldsymbol{\epsilon}_{\text{obs}}$ and $\boldsymbol{\epsilon}_{\text{unobs}}$ are independent standard Gaussian tensors with the same shape as $\mathbf{q}_s$; equivalently, their vectorized forms are distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I}_{\text{lat}})$, where $\mathbf{I}_{\text{lat}} \in \mathbb{R}^{D\times D}$ with $D = c_s 16^3$ is the identity covariance over the full latent state, and $\odot$ denotes element-wise multiplication. For observed cells, increasing C_i simultaneously strengthens the retained source contribution and suppresses the stochastic component, whereas unsupported cells carry no source-dependent mean and reduce to the Gaussian prior.

Induced endpoint entropy. The above parameterization makes the reliability–uncertainty relation explicit. Conditioned on $(\mathbf{q}_{s,i}, M_i, C_i)$, each spatial cell i of the 16^3 latent grid satisfies

$$\tilde{\mathbf{q}}_{s,i} \sim \begin{cases} \mathcal{N}(C_i \mathbf{q}_{s,i}, (1-C_i)^2 \mathbf{I}_{c_s}), & M_i = 1, \\ \mathcal{N}(\mathbf{0}, \mathbf{I}_{c_s}), & M_i = 0, \end{cases} \qquad h_i(C_i) = \frac{c_s}{2} \log\big(2\pi e (1-C_i)^2\big) \quad (M_i = 1, C_i < 1), \tag{7}$$

where $\mathbf{q}_{s,i}, \tilde{\mathbf{q}}_{s,i} \in \mathbb{R}^{c_s}$ denote the source and stochastic source latent vectors at cell i, $\mathbf{I}_{c_s}$ is the $c_s \times c_s$ identity matrix, and $h_i(C_i)$ denotes the conditional differential entropy of the stochastic source latent at an observed cell.

Because $\partial h_i / \partial C_i = -c_s/(1-C_i) < 0$, confidence induces a monotonic ordering of local source-endpoint uncertainty: high-confidence cells approach deterministic source anchors, uncertain observed cells retain greater stochastic freedom for correction, and unsupported cells revert to the Gaussian prior at the source boundary. This establishes the entropy-structured interpretation of $p_s^{\mathbf{M},\mathbf{C}}$, which requires only that $\mathbf{C}$ provide a monotonic reliability signal, not a calibrated posterior probability.

Endpoint entropy is distinct from the path-space relative entropy in equation 2: the former structures uncertainty at the source boundary, whereas the latter regularizes stochastic transport between endpoints. Unlike variance-only source modeling, ERSB couples source retention with confidence-dependent uncertainty, preserving reliable evidence while allocating greater corrective freedom to less reliable regions; the variance-only alternative is evaluated in Appendix E.

3.3 Reliability-Guided reverse Repair Dynamics

ERSB structures source uncertainty but cannot enforce reliability-aware reverse repair. RGRD addresses this through region-aware supervision for preservation, correction, and completion, and RLGA for reliability-guided information propagation and token updates.

Region-aware supervision. To separate preservation, correction, and completion, for latent tensors $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{c_s \times 16\times 16\times 16}$ and a single-channel nonnegative spatial weight map $\mathbf{W}$, we define

$$\mathcal{D}(\mathbf{A}, \mathbf{B}; \mathbf{W}) = \frac{\big\|\mathbf{W}^{1/2} \odot (\mathbf{A} - \mathbf{B})\big\|_2^2}{\max(c_s \|\mathbf{W}\|_1, 10^{-6})},$$

where $\mathbf{W}^{1/2}$ denotes the element-wise square root of the spatial weight map, $\mathbf{W}$ is broadcast over the latent-channel dimension, $\|\mathbf{W}\|_1$ denotes the sum of its spatial weights, $\|\cdot\|_2$ denotes the vectorized Euclidean norm over all latent entries, and 10^{-6} is a numerical stabilizer. The region-specific objectives are

$$\mathcal{L}_{\text{high}} = \mathcal{D}(\hat{\mathbf{q}}_g, \mathbf{q}_s; \mathbf{M} \odot \mathbf{C}), \quad \mathcal{L}_{\text{low}} = \mathcal{D}(\hat{\mathbf{q}}_g, \mathbf{q}_g; \mathbf{M} \odot (1-\mathbf{C})), \quad \mathcal{L}_{\text{unobs}} = \mathcal{D}(\hat{\mathbf{q}}_g, \mathbf{q}_g; 1-\mathbf{M}). \tag{8}$$

High-confidence source-supported regions are anchored to $\mathbf{q}_s$, while uncertain and unsupported regions are supervised toward $\mathbf{q}_g$. These task-specific endpoint regularizers do not modify the formal paired path-space KL projection or the reference conditional bridge law; instead, they intentionally shape the learned endpoint estimator toward source preservation, uncertain-region correction, and unsupported-region completion. Here, $\mathcal{L}_{\text{cons}}$ summarizes the weighted region-specific objectives.

We additionally match latent gradients near reliability and support transitions using the boundary-continuity loss $\mathcal{L}_{\text{bd}}$, whose construction is detailed in Appendix B. The total objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{bridge}} \mathcal{L}_{\text{bridge}} + \lambda_{\text{high}} \mathcal{L}_{\text{high}} + \lambda_{\text{low}} \mathcal{L}_{\text{low}} + \lambda_{\text{unobs}} \mathcal{L}_{\text{unobs}} + \lambda_{\text{bd}} \mathcal{L}_{\text{bd}}, \tag{9}$$

where λ_{bridge}, λ_{high}, λ_{low}, λ_{unobs}, and λ_{bd} are scalar weights controlling bridge regression, reliable-region preservation, uncertain-region correction, unsupported-region completion, and boundary-continuity supervision, respectively.

Reliability-aware local geometric attention. The regional objective specifies the desired preservation/correction behavior, yet it does not by itself impose reliability-asymmetric information flow inside the repair network. Standard self-attention does not explicitly enforce such asymmetry or local 3D geometric structure, even when confidence is available in the token features. RLGA therefore restricts attention to fixed local 3D neighborhoods and uses reliability together with local geometry to direct information propagation during repair.

For latent-grid locations i and j, let $\mathbf{p}_i, \mathbf{p}_j \in \mathbb{R}^3$ denote their 3D grid coordinates and let $\hat{\mathbf{n}}_i, \hat{\mathbf{n}}_j \in \mathbb{R}^3$ denote source-derived unit orientation cues. We construct

$$\mathbf{r}_{ij} = [\Delta\mathbf{p}_{ij}, d_{ij}, s^n_{ij}] \in \mathbb{R}^5, \qquad \mathbf{b}^{\mathrm{geo}}_{ij} = \phi_g(\mathbf{r}_{ij}) \in \mathbb{R}^H, \tag{10}$$

where $\Delta\mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i$, $d_{ij} = \|\Delta\mathbf{p}_{ij}\|_2$, and $s^n_{ij} = \hat{\mathbf{n}}_i^\top \hat{\mathbf{n}}_j$. Here, $\phi_g(\cdot)$ is a lightweight MLP that maps the 5-D geometric relation descriptor to H head-specific attention biases, with $H = 8$ denoting the number of attention heads. The orientation cues are derived only from the source geometry and use no ground-truth information; when a valid cue is unavailable, s^n_{ij} is set to zero.

Let $\mathbf{X}^{(l)} = \{\mathbf{x}^{(l)}_i\}_{i=1}^{N_{\mathrm{tok}}}$ denote the token features at the l-th RLGA block, where $N_{\mathrm{tok}} = 16^3$ is the number of latent-grid tokens. Each block uses a token embedding dimension $d_{\mathrm{model}} = 256$, yielding a per-head feature dimension $d_h = d_{\mathrm{model}}/H = 32$. The query and key features $\mathbf{q}^{(l,h)}_i, \mathbf{k}^{(l,h)}_j \in \mathbb{R}^{d_h}$ are obtained from learnable linear projections of $\mathbf{X}^{(l)}$. Let $\mathcal{N}(i)$ denote the fixed local 3D neighborhood of token i. For attention head $h \in \{1, \ldots, H\}$, the corresponding logit is

$$e^{(l,h)}_{ij} = \frac{(\mathbf{q}^{(l,h)}_i)^\top \mathbf{k}^{(l,h)}_j}{\sqrt{d_h}} + \lambda^{(l)}_g b^{\mathrm{geo},h}_{ij} + \lambda^{(l)}_c \log\big(\max(C_j, C_{\min})\big), \quad j \in \mathcal{N}(i). \tag{11}$$

Here, $b^{\mathrm{geo},h}_{ij}$ denotes the h-th component of the geometric bias in equation 10, and C_j is the confidence of neighboring token j, which acts as an information provider. The coefficients $\lambda^{(l)}_g$ and $\lambda^{(l)}_c$ are learnable scalar parameters independently maintained by each RLGA block and shared across its attention heads, while $C_{\min} = 0.05$ is a confidence floor used for numerical stability.

The confidence term favors reliable neighboring tokens as information providers. If $\mathbf{u}^{(l)}_i$ denotes the aggregated multi-head attention output, the receiver-side residual update is $\Delta\mathbf{x}^{(l)}_i = (1 - C_i)\mathbf{u}^{(l)}_i$, so reliable tokens preferentially guide nearby regions while receiving smaller modifications themselves, whereas uncertain and unsupported tokens are allowed larger corrections. We use $K_{\mathrm{nbr}} = 16$ neighbors per token. Neighborhood construction, orientation preprocessing, learnable-scale initialization, and the complete Transformer block are given in Appendix C. Overall, RGRD preserves reliable structure while concentrating repair on uncertain and unsupported regions, with RLGA reducing attention complexity from $O(N^2_{\mathrm{tok}} d_{\mathrm{model}})$ to $O(N_{\mathrm{tok}} K_{\mathrm{nbr}} d_{\mathrm{model}})$.

4 Experiments

We evaluate Entro-Bridge on object- and scene-level single-view reconstruction. The object- and scene-level models are trained on Toys4K and HSSD, respectively, while ABO and 3D-FRONT are used exclusively for evaluation. Additional implementation details, rendering metrics, ablations, hyperparameter analyses, and efficiency measurements are provided in the appendices.

4.1 Experimental Setup

Object- and scene-level models are trained on Toys4K and HSSD and evaluated on ABO and 3D-FRONT, respectively, without target-set training or fine-tuning. VGGT and TRELLIS are frozen; only F_θ is optimized with AdamW for 30k iterations (learning rate 5×10^{-4}; batch size 1). Inference uses five deterministic reverse-bridge steps.

We report CD and F-score for geometry, and PSNR-N and LPIPS-N from normal maps rendered at identical unseen poses. Preprocessing, architecture, training, decoding, and evaluation details are provided in Appendices B and C; full RGB results appear in Appendix D.

4.2 Main Comparisons

We compare Entro-Bridge with representative reconstruction and generation baselines. External methods use their released pretrained checkpoints and native pipelines without additional fine-tuning on ABO or 3D-FRONT. All methods share the same test inputs, camera poses and metrics. Therefore, Table 1 reports end-to-end performance under a common evaluation protocol.

Table 1: **Quantitative comparisons on ABO and 3D-FRONT.** Mean $\pm$ standard deviation over three runs. The main text reports geometry and normal-rendering metrics, while complete RGB rendering metrics are provided in Appendix D.

Dataset	Method	CD ↓	F-score (%) ↑	PSNR-N (dB) ↑	LPIPS-N ↓
ABO	GaussianAnything (Lan et al., 2025)	0.083 ± 0.0014	51.70 ± 0.62	21.174 ± 0.374	0.201 ± 0.0048
	Real3D (Jiang et al., 2025)	0.066 ± 0.0013	57.00 ± 0.58	21.511 ± 0.361	0.179 ± 0.0045
	LGM (Tang et al., 2024)	0.074 ± 0.0012	49.20 ± 0.65	20.228 ± 0.326	0.205 ± 0.0049
	VoxHammer (Li et al., 2026)	0.047 ± 0.0011	73.10 ± 0.51	20.476 ± 0.339	0.160 ± 0.0042
	TRELLIS (Xiang et al., 2025)	0.035 ± 0.0011	84.00 ± 0.43	24.072 ± 0.388	0.106 ± 0.0037
	SAM3D (Chen et al., 2026)	0.032 ± 0.0012	84.40 ± 0.41	24.115 ± 0.392	0.102 ± 0.0035
	BridgeShape (Kong et al., 2026)	0.028 ± 0.0012	86.00 ± 0.39	24.650 ± 0.380	0.120 ± 0.0037
	† Points-to-3D (Xia et al., 2026)	0.023 ± 0.0011	89.00 ± 0.36	24.920 ± 0.397	0.086 ± 0.0032
	Ours	$\mathbf{0.0200 \pm 0.0004}$	$\mathbf{92.00 \pm 0.28}$	$\mathbf{26.890 \pm 0.110}$	$\mathbf{0.0760 \pm 0.0009}$
3D-FRONT	TouchDream (Wang et al., 2026)	0.044 ± 0.0014	48.70 ± 0.72	19.108 ± 0.368	0.263 ± 0.0049
	TRELLIS (Xiang et al., 2025)	0.095 ± 0.0015	47.10 ± 0.61	18.512 ± 0.324	0.253 ± 0.0047
	SceneGen (Meng et al., 2026)	0.087 ± 0.0013	49.40 ± 0.68	19.429 ± 0.379	0.233 ± 0.0045
	MIDI (Huang et al., 2025)	0.076 ± 0.0012	52.10 ± 0.64	21.187 ± 0.386	0.167 ± 0.0036
	† Points-to-3D (Xia et al., 2026)	0.039 ± 0.0011	75.50 ± 0.52	21.348 ± 0.394	0.163 ± 0.0033
	VoxHammer (Li et al., 2026)	0.052 ± 0.0010	69.70 ± 0.55	20.770 ± 0.381	0.184 ± 0.0038
	Ours	$\mathbf{0.035 \pm 0.0010}$	$\mathbf{81.30 \pm 0.36}$	$\mathbf{22.980 \pm 0.397}$	$\mathbf{0.152 \pm 0.0031}$

† Points-to-3D is reimplemented and trained by us following the published method; results may differ from an official implementation.

As shown in Table 1, Entro-Bridge consistently outperforms the strongest baselines on both ABO and 3D-FRONT, reducing CD by approximately 13.0% and 10.3% relative to Points-to-3D, with F-scores of 92.0% and 81.3%, respectively. The gains in PSNR-N and LPIPS-N further indicate improved local surface orientation and structural continuity. Figure 3 presents qualitative comparisons from the same unseen camera poses. The source reconstructions contain gaps in occluded or back-facing regions, while several comparison methods retain boundary discontinuities or overly smoothed local structures. Entro-Bridge more closely follows the reference geometry while preserving the recognizable source-supported regions.

4.3 Ablation and Efficiency

We conduct controlled ablations on ABO using identical source reconstructions, training settings, and frozen TRELLIS components. We evaluate ERSB, RGRD, RLGA, and the paired-bridge backbone; for *w/o RLGA*, RLGA is replaced with standard self-attention.

Table 2 shows consistent degradation when removing the reliability-aware components, while noise-start conditional diffusion causes the largest drop. The finer ablations further support the proposed ERSB formulation and the reliability- and geometry-aware design of RGRD.

Table 2: **Core ablations on ABO.** Results are reported as mean $\pm$ standard deviation over three independent training runs. Complete eight-metric results and additional fine-grained ablations are provided in Appendix E.

Variant	CD ↓	F-score (%) ↑	PSNR-N (dB) ↑	LPIPS-N ↓
Full Entro-Bridge	$\mathbf{0.0200 \pm 0.0004}$	$\mathbf{92.0 \pm 0.28}$	$\mathbf{26.89 \pm 0.11}$	$\mathbf{0.0760 \pm 0.0009}$
Vanilla Paired SB	0.0310 ± 0.0007	83.4 ± 0.52	24.05 ± 0.18	0.1390 ± 0.0018
w/o ERSB	0.0230 ± 0.0005	88.5 ± 0.41	25.48 ± 0.14	0.0970 ± 0.0014
w/o Observation Mask $\mathbf{M}$	0.0210 ± 0.0006	90.1 ± 0.44	26.01 ± 0.15	0.0850 ± 0.0015
w/o Confidence $\mathbf{C}$	0.0210 ± 0.0004	91.2 ± 0.36	26.35 ± 0.12	0.0780 ± 0.0012
Variance-Level Endpoint Mixing	0.0207 ± 0.0005	90.6 ± 0.45	26.18 ± 0.16	0.0840 ± 0.0016
w/o Boundary-Continuity Loss	0.0200 ± 0.0004	90.8 ± 0.34	26.14 ± 0.13	0.0830 ± 0.0011
w/o RLGA	0.0260 ± 0.0008	86.0 ± 0.63	24.72 ± 0.20	0.1180 ± 0.0021
w/o RLGA Confidence Bias	0.0200 ± 0.0005	90.9 ± 0.39	26.31 ± 0.14	0.0790 ± 0.0013
w/o RLGA Geometry Bias	0.0220 ± 0.0006	89.1 ± 0.51	25.67 ± 0.17	0.0920 ± 0.0017
Bridge $\rightarrow$ Noise-Start Conditional Diffusion	0.0390 ± 0.0009	82.3 ± 0.74	23.95 ± 0.24	0.1460 ± 0.0024

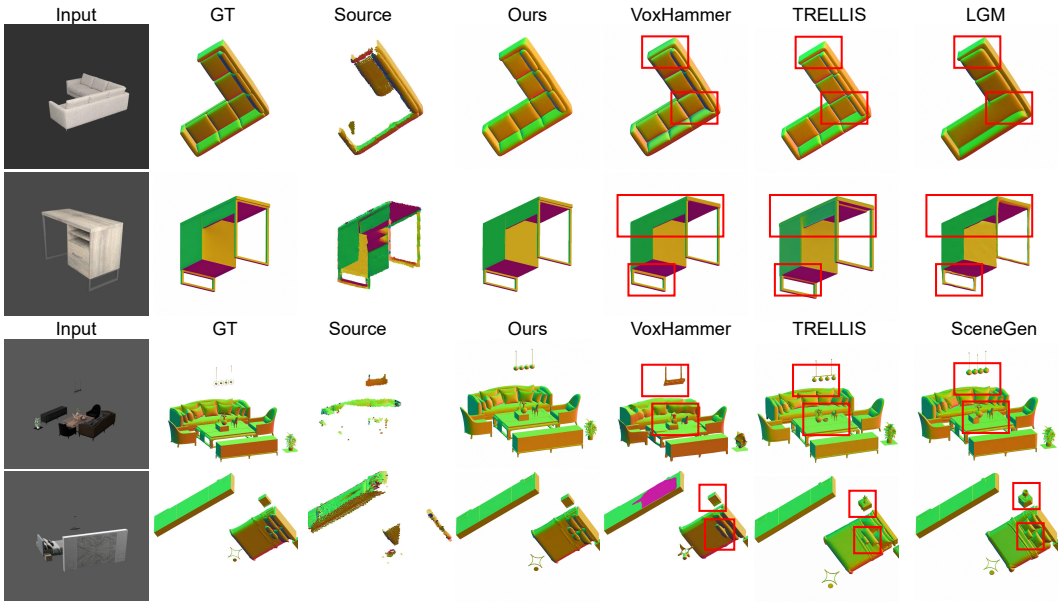

Figure 3: Qualitative comparison on ABO and 3D-FRONT datasets.

To examine the completion preservation trade-off beyond global metrics, we further evaluate GT-defined visible and unseen regions independently of the internal reliability masks.As shown in Table 3, Entro-Bridge improves unseen-region completion while achieving the lowest visible-region error and Anchor Drift, supporting its completion preservation balance. Detailed definitions are provided in Appendix E.4.

Table 3: **Region-wise completion and preservation analysis on ABO.** Results are reported as mean $\pm$ standard deviation over three independent training runs. Unseen metrics evaluate completion, Observed GT Error measures visible-region accuracy, and Anchor Drift quantifies reliable-source preservation.

Method	Unseen CD ↓	Unseen F-score (%) ↑	Obs. GT Error ↓	Anchor Drift ↓
† Points-to-3D	0.0240 ± 0.0006	81.2 ± 0.46	0.0094 ± 0.0003	N/A*
Vanilla Paired SB	0.0310 ± 0.0008	77.5 ± 0.64	0.0108 ± 0.0005	0.0101 ± 0.0005
w/o Confidence-Aware Source	0.0260 ± 0.0007	83.6 ± 0.57	0.0082 ± 0.0004	0.0084 ± 0.0004
Full Entro-Bridge	$\mathbf{0.0210 \pm 0.0005}$	$\mathbf{87.4 \pm 0.38}$	$\mathbf{0.0070 \pm 0.0002}$	$\mathbf{0.0073 \pm 0.0003}$

† Points-to-3D is reimplemented following the published method; results may differ from an official implementation.

* Anchor Drift is reported only for methods sharing the same VGGT source reconstruction.

In our implementation, the repair module contains 6.958M parameters and requires 0.606 s per sample when using five reverse steps. Complete analyses of computational efficiency and hyperparameter sensitivity are provided in Appendices G and F, respectively. .

5 Discussion and Conclusion

Entro-Bridge improves single-view 3D reconstruction by explicitly modeling the spatially varying reliability of feed-forward geometry. Experiments on ABO and 3D-FRONT show consistent gains in reconstruction quality while preserving reliable source structures, and the ablations confirm the complementary roles of ERSB and reliability-guided reverse repair. These results support the core design principle of Entro-Bridge: reliable geometry should be preserved, uncertain predictions should be corrected, and unsupported regions should be completed within a unified reconstruction process. The current framework still depends on the quality of upstream confidence estimation and the frozen TRELLIS latent representation. More accurate reliability estimation, higher-resolution latent spaces, and extension to other reconstruction backbones provide promising directions for future work.

AI USE STATEMENT

We used generative AI tools to assist with the translation and language editing of portions of the manuscript, including improving grammar, clarity, and readability. We did not use generative AI tools to generate synthetic datasets, formulate mathematical claims or proofs, or conduct qualitative data analysis. The research methodology, experimental results, and scientific claims presented in this work were developed and verified by the authors. All AI-assisted text was carefully reviewed and revised by the authors. We take full responsibility for the final content of this work, including any text produced or edited with the assistance of generative AI.

ETHICS STATEMENT

We confirm adherence to the ICLR Code of Ethics and have carefully considered the ethical implications of our research. Our work focuses on sparse-view 3D scene reconstruction and novel-view synthesis using established publicly available benchmark datasets. We do not collect new data involving human subjects or process personally identifiable information. The proposed method is intended for research purposes, and its use should comply with the licenses and usage policies of the corresponding datasets.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed framework, optimization objectives, and training procedure in the main paper. The experimental section specifies the evaluation metrics, implementation settings, and training configurations used in our experiments. Detailed implementation specifics are provided in Appendices A, B and C, while additional ablation studies and hyperparameter analyses are presented in Appendices E and F, respectively.

REFERENCES

Yuanhao Cai, He Zhang, Kai Zhang, Yixun Liang, Mengwei Ren, Fujun Luan, Qing Liu, Soo Ye Kim, Jianming Zhang, Zhifei Zhang, Yuqian Zhou, Yulun Zhang, Xiaokang Yang, Zhe Lin, and Alan Yuille. Baking gaussian splatting into diffusion denoiser for fast and scalable single-stage image-to-3D generation and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25062–25072, 2025.

Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. Reconviagen: Towards accurate multi-view 3D object reconstruction via generation. In *International Conference on Learning Representations*, 2026.

Xingyu Chen, FU-JEN CHU, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, June 2026.

Ruihang Chu, Enze Xie, Shentong Mo, Zhenguo Li, Matthias Nießner, Chi-Wing Fu, and Jiaya Jia. DiffComplete: Diffusion-based generative 3D shape completion. *Advances in Neural Information Processing Systems*, 36:75951–75966, 2023.

Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3D-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5868–5877, 2017.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709, 2021.

Yitong Dong, Qi Zhang, Minchao Jiang, Zhiqiang Wu, Qingnan Fan, Ying Feng, Huaqi Zhang, Hujun Bao, and Guofeng Zhang. One-shot refiner: Boosting feed-forward novel view synthesis via one-step diffusion. volume 40, pp. 3659–3667, 2026.

Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE international conference on computer vision*, pp. 85–93, 2017.

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.

Yuzhou Ji, Qijian Tian, He Zhu, Xiaoqi Jiang, Guangzhi Cao, Lizhuang Ma, Yuan Xie, and Xin Tan. S2D: Sparse to dense lifting for 3D reconstruction with minimal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14219–14228, 2026.

Hanwen Jiang, Qixing Huang, and Georgios Pavlakos. Real3D: Scaling up large reconstruction models with real-world images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5821–5833, 2025.

Xuancheng Jin, Rengan Xie, Wenting Zheng, Rui Wang, Hujun Bao, and Yuchi Huo. Fuse3D: Generating 3D assets controlled by multi-image fusion. In *ACM SIGGRAPH Asia Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia*, pp. 144:1–144:12, 2025.

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

Dequan Kong, Honghua Chen, Zhe Zhu, and Mingqiang Wei. BridgeShape: Latent diffusion schrödinger bridge for 3D shape completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5726–5734, 2026.

Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. GaussianAnything: Interactive point cloud latent diffusion for 3D generation. In *International Conference on Learning Representations*, pp. 1–10, 2025.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with mast3r. In *Proceedings of the European Conference on Computer Vision*, volume 15130 of *Lecture Notes in Computer Science*, pp. 71–91, 2024.

Lin Li, Zehuan Huang, Haoran Feng, Gengxiong Zhuang, Rui Chen, Chunchao Guo, and Lu Sheng. VoxHammer: Training-free precise and coherent 3D editing in native 3D space. In *International Conference on 3D Vision*, pp. 1281–1292, 2026.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 2023.

Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. SceneGen: Single-image 3D scene generation in one feedforward pass. In *International Conference on 3D Vision 2026*, pp. 543–553, 2026.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. AutoSDF: Shape priors for 3D completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 306–315, 2022.

Seungtae Nam, Xiangyu Sun, Gyeongjin Kang, Younggeun Lee, Seungjun Oh, and Eunbyung Park. Generative densification: Learning to densify gaussians for high-fidelity generalizable 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26683–26693, 2025.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Chen Shi, Shaoshuai Shi, Xiaoyang Lyu, Chunyang Liu, Kehua Sheng, Bo Zhang, and Li Jiang. UniSplat: Unified spatio-temporal fusion via 3D latent scaffolds for dynamic driving scene reconstruction. In *International Conference on Learning Representations*, pp. 1–10, 2026.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3D content creation. In *Proceedings of the European Conference on Computer Vision*, volume 15062 of *Lecture Notes in Computer Science*, pp. 1–18, 2024.

Shengji Tang, Weicai Ye, Peng Ye, Weihao Lin, Yang Zhou, Tao Chen, and Wanli Ouyang. HiSplat: Hierarchical 3D gaussian splatting for generalizable sparse-view reconstruction. In *International Conference on Learning Representations*, pp. 1–10, 2025.

Yan Tian, Pengcheng Xue, Weiping Ding, Mahmoud Hassaballah, Karen Egiazarian, Aura Conci, Abdulkadir Sengur, and Leszek Rutkowski. DM-CFO: A diffusion model for compositional 3D tooth generation with collision-free optimization. *IEEE Transactions on Visualization and Computer Graphics*, 32(7):5470–5482, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Yuanbo Wang, Xinning Wang, Zhaoxuan Zhang, Changlong Wang, Qianchen Xia, Xiaopeng Wei, and Xin Yang. TouchDream: 3D object completion through imagined touch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8901–8910, June 2026.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3D+: Improving 3D reconstructions with single-step diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26024–26035, 2025a.

Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal shape completion via conditional generative adversarial networks. In *Proceedings of the European Conference on Computer Vision*, volume 12349 of *Lecture Notes in Computer Science*, pp. 281–296, 2020.

Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3D reconstruction from occluded 2d images. In *2025 IEEE/CVF International Conference on Computer Vision*, pp. 9181–9193. IEEE, 2025b.

Jiatong Xia, Zicheng Duan, Anton van den Hengel, and Lingqiao Liu. Points-to-3D: Structure-aware 3D generation with point cloud priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19928–19939, June 2026.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025.

Pengcheng Xue, Yan Tian, Qiutao Song, Ziyi Wang, Linyang He, Weiping Ding, Mahmoud Hassaballah, Karen Egiazarian, Wei-fa Yang, and Leszek Rutkowski. SVGS: Single-view to 3D object editing via gaussian splatting. *ACM Transactions on Multimedia Computing Communications and Applications*, 22(8):1–19, 2026.

Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6239–6249, 2022.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision*, volume 11212 of *Lecture Notes in Computer Science*, pp. 785–801, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5525–5534, 2019.

Botao Ye, Sifei Liu, Haofei Xu, Li Xueting, Marc Pollefeys, Ming-Hsuan Yang, and Peng Songyou. No pose, no problem: Surprisingly simple 3D gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, pp. 1–10, 2025.

Zehao Yu and Shenghua Gao. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1946–1955, 2020.

Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3D gaussian splatting from sparse multi-view images with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9869–9877, 2025.

Xinxin Zhao, Jian Jiang, Yan Tian, Liqin Wu, Zhaocheng Xu, Wei-fa Yang, Yunuo Zou, and Xun Wang. Innovative tooth segmentation using hierarchical features and bidirectional sequence modeling. *Pattern Recognition*, 175:113045, 2026a.

Xinxin Zhao, Jinpeng Ye, Bo Wei, Liqin Wu, Mahmoud Hassaballah, Karen Egiazarian, Aura Conci, Victor Hugo C. de Albuquerque, Abdulkadir Sengur, Leszek Rutkowski, and Yan Tian. Fumamba: A frequency-enhanced dynamic scanning framework for oralscan image segmentation. *Neurocomputing*, 701:134618, 2026b.

A Bridge Formulation and Reverse Sampling

This appendix provides the path-space KL decomposition underlying the paired formulation, the complete conditional bridge coefficients, variance schedule, residual-learning interpretation, and deterministic reverse-bridge sampling equations omitted from the main text for concision.

A.1 Paired Entropy Projection and Conditional Bridge

Let $\mathbb{R}_{\mathrm{ref}}$ denote the drift-free Gaussian reference path measure and let $\pi_{\mathrm{ref}}(\mathbf{q}_0, \mathbf{q}_T)$ be its joint endpoint law. For any candidate path measure $\mathbb{P}$, denote its endpoint coupling by $\pi_{\mathbb{P}}(\mathbf{q}_0, \mathbf{q}_T)$. Under the standard absolute-continuity conditions for the corresponding Radon–Nikodym derivatives, disintegrating both path measures with respect to their endpoints gives the KL chain rule

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{R}_{\mathrm{ref}}) = D_{\mathrm{KL}}(\pi_{\mathbb{P}}\|\pi_{\mathrm{ref}}) \\ + \mathbb{E}_{(\mathbf{q}_0,\mathbf{q}_T)\sim\pi_{\mathbb{P}}}\left[D_{\mathrm{KL}}(\mathbb{P}(\cdot \mid \mathbf{q}_0, \mathbf{q}_T) \,\|\, \mathbb{R}_{\mathrm{ref}}(\cdot \mid \mathbf{q}_0, \mathbf{q}_T))\right]. \end{aligned} \tag{12}$$

The generic Schrödinger Bridge in equation 2 fixes only the two endpoint marginals, so the joint endpoint coupling is part of the entropy-regularized transport problem. In our supervised setting, each training instance supplies a paired complete endpoint and feed-forward source endpoint. Together with the stochastic source construction in equation 6, these pairs induce an empirical joint coupling $\pi_{\mathrm{pair}}(\mathbf{q}_0, \mathbf{q}_T)$. Accordingly, the paired path-space problem underlying the bridge regression component is

$$\mathbb{P}^{\star}_{\mathrm{pair}} = \arg\min_{\mathbb{P}:\, \pi_{\mathbb{P}}=\pi_{\mathrm{pair}}} D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{R}_{\mathrm{ref}}), \tag{13}$$

which uses the stronger paired boundary condition

$$\pi_{\mathbb{P}}(\mathbf{q}_0, \mathbf{q}_T) = \pi_{\mathrm{pair}}(\mathbf{q}_0, \mathbf{q}_T), \qquad \mathbf{q}_0 = \mathbf{q}_g, \quad \mathbf{q}_T = \tilde{\mathbf{q}}_s. \tag{14}$$

Under this constraint, the first term of equation 12 is constant. The conditional KL terms are nonnegative and attain their minimum at

$$\mathbb{P}^{\star}_{\mathrm{pair}}(\cdot \mid \mathbf{q}_0, \mathbf{q}_T) = \mathbb{R}_{\mathrm{ref}}(\cdot \mid \mathbf{q}_0, \mathbf{q}_T). \tag{15}$$

Equivalently, the paired entropy projection is the mixture

$$\mathbb{P}^{\star}_{\mathrm{pair}} = \int \mathbb{R}_{\mathrm{ref}}(\cdot \mid \mathbf{q}_0, \mathbf{q}_T)\, \pi_{\mathrm{pair}}(d\mathbf{q}_0, d\mathbf{q}_T). \tag{16}$$

Thus, paired supervision fixes the endpoint coupling, and the remaining path-space entropy minimization is solved conditionally by the pinned reference bridges. This is the precise paired-data specialization used by Entro-Bridge: it retains the entropy-projection interpretation of the Schrödinger formulation while avoiding numerical estimation of an unknown endpoint coupling or coupled Schrödinger potentials.

For the drift-free Gaussian reference process, the pinned bridge has a Gaussian time marginal

$$q(\mathbf{q}_t \mid \mathbf{q}_0, \mathbf{q}_T) = \mathcal{N}\left(\mathbf{q}_t; \boldsymbol{\mu}_t(\mathbf{q}_0, \mathbf{q}_T), \boldsymbol{\Sigma}_t\right), \tag{17}$$

which is the tractable conditional bridge sampled during training. The semantic task direction and diffusion-time convention should be distinguished: Entro-Bridge restores $\mathbf{q}_T = \tilde{\mathbf{q}}_s$ into $\mathbf{q}_0 = \mathbf{q}_g$, whereas the reference bridge is indexed from complete geometry at $t = 0$ to the incomplete source at $t = T$.

A.2 Variance Schedule and Bridge Coefficients

For a drift-free reference diffusion with the discrete variance schedule $\{\beta_t\}_{t=1}^{T}$, the accumulated variances measured from the two endpoints are

$$\begin{aligned} \sigma_t^2 &= \sum_{\tau=1}^{t} \beta_\tau, \\ \sigma_{b,t}^2 &= \sum_{\tau=t+1}^{T} \beta_\tau. \end{aligned} \tag{18}$$

We use $T = 1000$ and the implemented variance schedule

$$\begin{aligned} \beta_t &= \left[\sqrt{\beta_{\min}} + \frac{t-1}{T-1}\left(\sqrt{\beta_{\max}} - \sqrt{\beta_{\min}}\right)\right]^2, \\ \beta_{\min} &= 10^{-4}, \qquad \beta_{\max} = 2 \times 10^{-2}. \end{aligned} \tag{19}$$

The conditional bridge mean is

$$\boldsymbol{\mu}_t(\mathbf{q}_0, \mathbf{q}_T) = \mu_{0,t}\mathbf{q}_0 + \mu_{1,t}\mathbf{q}_T, \tag{20}$$

where

$$\mu_{0,t} = \frac{\sigma_{b,t}^2}{\sigma_t^2 + \sigma_{b,t}^2}, \qquad \mu_{1,t} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{b,t}^2}. \tag{21}$$

The conditional covariance is

$$\boldsymbol{\Sigma}_t = \sigma_{\mathrm{br},t}^2 \mathbf{I}, \qquad \sigma_{\mathrm{br},t}^2 = \frac{\sigma_t^2 \sigma_{b,t}^2}{\sigma_t^2 + \sigma_{b,t}^2}. \tag{22}$$

Therefore, a bridge state can be sampled as

$$\mathbf{q}_t = \mu_{0,t}\mathbf{q}_0 + \mu_{1,t}\mathbf{q}_T + \sigma_{\mathrm{br},t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{23}$$

This parameterization satisfies

$$\mathbf{q}_t \to \mathbf{q}_0 \quad \text{as} \quad t \to 0, \qquad \mathbf{q}_t \to \mathbf{q}_T \quad \text{as} \quad t \to T, \tag{24}$$

which makes the roles of the two endpoints explicit.

A.3 Residual Regression as Reverse Endpoint Estimation

Under the conditional bridge sample in equation 23, the training target is

$$\mathbf{v}_t = \frac{\mathbf{q}_t - \mathbf{q}_0}{\sigma_t}, \qquad \hat{\mathbf{q}}_0 = \mathbf{q}_t - \sigma_t F_\theta(\mathbf{q}_t, t, \mathbf{q}_T, \mathbf{M}, \mathbf{C}). \tag{25}$$

Therefore, for every sampled bridge state,

$$\|F_\theta(\mathbf{q}_t, t, \mathbf{q}_T, \mathbf{M}, \mathbf{C}) - \mathbf{v}_t\|_2^2 = \frac{1}{\sigma_t^2}\,\|\hat{\mathbf{q}}_0 - \mathbf{q}_0\|_2^2\,. \tag{26}$$

The residual loss is thus a time-weighted regression of the complete endpoint under samples from the paired conditional bridge. With squared loss, its population optimum is

$$F_\theta^\star(\mathbf{q}_t, t, \mathbf{q}_T, \mathbf{M}, \mathbf{C}) = \mathbb{E}\left[\frac{\mathbf{q}_t - \mathbf{q}_0}{\sigma_t}\,\middle|\,\mathbf{q}_t, \mathbf{q}_T, \mathbf{M}, \mathbf{C}, t\right], \tag{27}$$

which implies

$$\mathbf{q}_t - \sigma_t F_\theta^\star = \mathbb{E}[\mathbf{q}_0 \mid \mathbf{q}_t, \mathbf{q}_T, \mathbf{M}, \mathbf{C}, t]. \tag{28}$$

equation 12–equation 28 make explicit the relation between the path-space formulation and the implemented bridge-regression objective: the KL objective determines the paired conditional bridge law, while residual regression learns the reverse endpoint estimator associated with that law. The bridge regression term is therefore not presented as a direct numerical optimizer of the unrestricted marginal Schrödinger problem. The additional region-aware terms leave the reference conditional bridge law unchanged and instead shape the learned endpoint estimator toward task-specific preservation, correction, and completion.

A.4 Deterministic Reverse-Bridge Sampling

At inference time, the complete target geometry is unavailable. We construct $\tilde{\mathbf{q}}_s$ from the feed-forward reconstruction and initialize the reverse process from the incomplete source endpoint:

$$\mathbf{q}^{(K)} = \mathbf{q}_T = \tilde{\mathbf{q}}_s, \qquad T = t_K > t_{K-1} > \cdots > t_0 = 0. \tag{29}$$

At reverse step k, the network predicts

$$\hat{\mathbf{v}}_k = F_\theta\left(\mathbf{q}^{(k)}, t_k, \mathbf{q}_T, \mathbf{M}, \mathbf{C}\right) \tag{30}$$

and estimates the complete endpoint as

$$\hat{\mathbf{q}}_0^{(k)} = \mathbf{q}^{(k)} - \sigma_{t_k}\hat{\mathbf{v}}_k. \tag{31}$$

For efficient few-step inference, we use deterministic DDIM-style updates constructed from the learned endpoint estimator and the conditional-bridge coefficients. This sampler is an inference-time numerical trajectory; the stochastic conditional bridge above defines the training law and entropy projection. We first recover the noise direction associated with the current bridge state:

$$\hat{\boldsymbol{\epsilon}}_k = \frac{\mathbf{q}^{(k)} - \mu_{0,t_k}\hat{\mathbf{q}}_0^{(k)} - \mu_{1,t_k}\mathbf{q}_T}{\sigma_{\mathrm{br},t_k} + \varepsilon_{\mathrm{sam}}}, \tag{32}$$

where $\varepsilon_{\mathrm{sam}} > 0$ is a small numerical stabilizer.

The next reverse state is

$$\mathbf{q}^{(k-1)} = \mu_{0,t_{k-1}}\hat{\mathbf{q}}_0^{(k)} + \mu_{1,t_{k-1}}\mathbf{q}_T + \sigma_{\mathrm{br},t_{k-1}}\hat{\boldsymbol{\epsilon}}_k. \tag{33}$$

No fresh random noise is introduced during this update. Therefore, equation 33 should be understood as a deterministic discretization for efficient inference rather than as a separate solution of the path-space KL problem. The sampler proceeds from $t = T$ to $t = 0$. All main experiments use $K = 5$ reverse steps, and the final reconstructed latent is

$$\hat{\mathbf{q}}_g = \mathbf{q}^{(0)}. \tag{34}$$

The frozen TRELLIS sparse-structure decoder converts $\hat{\mathbf{q}}_g$ into the repaired active-voxel structure. For final asset reconstruction, the repaired structure is passed to the frozen TRELLIS Stage-2 structured-latent generator and the corresponding frozen output decoder. These downstream modules are inherited from TRELLIS and are not optimized as part of Entro-Bridge.

B Confidence Construction, Endpoint Entropy, and Region-Aware Supervision

This appendix describes how the observation-support map and confidence field are constructed, derives the local conditional entropy induced by the stochastic source endpoint, clarifies its distinction from path-space entropy regularization, and gives the complete boundary-weight formulation and loss configuration.

B.1 Source Geometry and Confidence Preprocessing

Given a single input image $\mathbf{I}$, we use a frozen VGGT model to predict depth, camera parameters, and the per-pixel depth confidence returned by its depth head. Each 3D point reconstructed by depth unprojection retains the confidence of its corresponding image pixel. Because the raw VGGT confidence is not restricted to $[0, 1]$, we first convert it into a relative reliability score. Specifically, nonnegative raw confidence values are rescaled by the sample-wise maximum whenever their maximum exceeds one, and are then mapped to $[0.05, 1]$:

$$c_n = 0.05 + 0.95\,\bar{c}_n, \qquad \bar{c}_n = \begin{cases} c_n^{\mathrm{raw}} / \max_m c_m^{\mathrm{raw}}, & \max_m c_m^{\mathrm{raw}} > 1, \\ c_n^{\mathrm{raw}}, & \text{otherwise.} \end{cases} \tag{35}$$

Thus, raw VGGT scores larger than one are normalized rather than directly clipped, preserving their relative ordering within each input.

The source points are transformed into the normalized coordinate system used by TRELLIS and voxelized onto a 64^3 grid. A binary occupancy indicator defines the observation-support grid. For every occupied voxel, we average the normalized confidence values of all points falling into that voxel; empty voxels are assigned zero confidence.

The support and confidence grids are then mapped to the 16^3 TRELLIS sparse-structure latent resolution using $4 \times 4 \times 4$ max pooling and average pooling, respectively. The confidence pooling is performed on the dense 64^3 grid, so empty child voxels contribute zero to the average. The resulting support mask $\mathbf{M}$ indicates whether a latent cell is supported by source geometry, while the continuous field $\mathbf{C}$ provides a normalized relative reliability signal for that support. Before being used by the source construction and repair network, we apply

$$\mathbf{C} \leftarrow \mathbf{M} \odot \mathrm{clip}(\mathbf{C}, 0, 1). \tag{36}$$

Consequently, unsupported latent locations always have zero confidence. Both $\mathbf{M}$ and $\mathbf{C}$ are precomputed and treated as fixed conditioning signals during bridge training. We use $\mathbf{C}$ as a relative reliability signal rather than as a calibrated posterior probability of geometric correctness.

B.2 Entropy-Structured Reliability-Aware Source Boundary

The confidence-aware endpoint is defined in equation 6 as

$$\tilde{\mathbf{q}}_s = \mathbf{M} \odot (\mathbf{C} \odot \mathbf{q}_s + (1 - \mathbf{C}) \odot \boldsymbol{\epsilon}_{\mathrm{obs}}) + (1 - \mathbf{M}) \odot \boldsymbol{\epsilon}_{\mathrm{unobs}}.$$

The support mask and confidence field have different functions:

- $\mathbf{M}$ identifies whether a latent location has source-geometry support.
- $\mathbf{C}$ measures the reliability of the available source geometry.
- $\mathbf{M} \odot \mathbf{C}$ identifies reliable source-supported regions.
- $\mathbf{M} \odot (1 - \mathbf{C})$ identifies low-confidence observed regions.
- $1 - \mathbf{M}$ identifies regions unsupported by the source reconstruction.

Conditioned on the source latent, support, and confidence, the noise injection gives a closed-form local terminal distribution. For latent cell i,

$$\tilde{\mathbf{q}}_{s,i} \mid (\mathbf{q}_{s,i}, M_i, C_i) \sim \begin{cases} \mathcal{N}\big(C_i \mathbf{q}_{s,i}, (1 - C_i)^2 \mathbf{I}_{c_s}\big), & M_i = 1, \\ \mathcal{N}(\mathbf{0}, \mathbf{I}_{c_s}), & M_i = 0. \end{cases} \tag{37}$$

For an observed cell with $C_i < 1$, the corresponding conditional differential entropy is

$$h_i(C_i) = \frac{c_s}{2} \log\big(2\pi e(1 - C_i)^2\big), \qquad \frac{\partial h_i}{\partial C_i} = -\frac{c_s}{1 - C_i} < 0. \tag{38}$$

Hence confidence induces a monotone spatial uncertainty field: higher confidence produces lower terminal entropy, and the limit $C_i \to 1$ becomes the deterministic anchor $\tilde{\mathbf{q}}_{s,i} = \mathbf{q}_{s,i}$. An unsupported cell remains $\mathcal{N}(\mathbf{0}, \mathbf{I}_{c_s})$. Because the injected Gaussian noise is independent across latent elements, equation 38 characterizes the local contributions to the conditional uncertainty of the source measure (with deterministic cells understood in the limiting sense). Differential entropy is used here only to characterize this monotone uncertainty structure in the fixed TRELLIS latent coordinates; it is not treated as an independently calibrated probability or as an additional training loss.

This endpoint entropy is conceptually different from the path-space relative entropy in equation 2. The endpoint term is induced by the confidence-dependent terminal distribution, whereas the path-space KL regularizes the stochastic transport law relative to the reference diffusion. Their coexistence gives Entro-Bridge two distinct entropic structures: where uncertainty is placed at the source boundary and how stochastic transport is regularized between endpoints.

We implement this uncertainty through amplitude-level source retention. Linear interpolation guarantees exact source preservation at $\mathbf{C} = 1$ and complete replacement by Gaussian noise at $\mathbf{C} = 0$; it jointly changes the retained mean $C_i \mathbf{q}_{s,i}$ and induces variance $(1 - C_i)^2$. This is different from directly assigning a confidence-dependent variance while leaving the source amplitude unchanged. The controlled variance-level alternative is evaluated in Appendix E.

B.3 Region-Aware Latent Supervision

For tensors $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{c_s \times 16 \times 16 \times 16}$ and a single-channel nonnegative spatial weight map $\mathbf{W}$, the masked latent distance is

$$\mathcal{D}(\mathbf{A}, \mathbf{B}; \mathbf{W}) = \frac{\big\|\mathbf{W}^{1/2} \odot (\mathbf{A} - \mathbf{B})\big\|_2^2}{\max\left(c_s \|\mathbf{W}\|_1, \varepsilon_{\text{loss}}\right)}, \qquad \varepsilon_{\text{loss}} = 10^{-6}. \tag{39}$$

Dividing by the weighted number of active latent elements makes the loss approximately invariant to region size and prevents spatial regions containing more cells from dominating the optimization.

Reliable observed regions are anchored to the source latent:

$$\mathcal{L}_{\text{high}} = \mathcal{D}\left(\hat{\mathbf{q}}_g, \mathbf{q}_s; \mathbf{M} \odot \mathbf{C}\right). \tag{40}$$

The source latent $\mathbf{q}_s$, rather than the complete target $\mathbf{q}_g$, is used as the reference in this region. This is intentional: the proposed framework treats high-confidence source-supported structures as reliable geometric anchors. Replacing $\mathbf{q}_s$ with $\mathbf{q}_g$ would encourage the network to rewrite source geometry even when it is considered reliable, which would conflict with the source-preserving objective.

Low-confidence observed regions are supervised toward the complete target:

$$\mathcal{L}_{\text{low}} = \mathcal{D}\left(\hat{\mathbf{q}}_g, \mathbf{q}_g; \mathbf{M} \odot (1 - \mathbf{C})\right). \tag{41}$$

Unsupported regions are also supervised toward the complete target:

$$\mathcal{L}_{\text{unobs}} = \mathcal{D}\left(\hat{\mathbf{q}}_g, \mathbf{q}_g; 1 - \mathbf{M}\right). \tag{42}$$

These terms use soft confidence weights rather than a hard confidence threshold. They therefore define a continuous transition among source preservation, selective correction, and generative completion.

B.4 Boundary-Continuity Supervision

To reduce discontinuities near observed–unobserved transitions and low-confidence boundaries, we additionally supervise local latent gradients.

We first compute local support and confidence statistics:

$$\begin{aligned}\bar{\mathbf{M}} &= \mathrm{AvgPool}_{3^3}(\mathbf{M}),\\ \bar{\mathbf{C}} &= \mathrm{clip}\left(\mathrm{AvgPool}_{3^3}(\mathbf{C}), 0, 1\right).\end{aligned} \tag{43}$$

The observed–unobserved transition weight is

$$\mathbf{B}_{\mathrm{tr}} = 4\ \mathrm{clip}\left(\bar{\mathbf{M}} \odot (1 - \bar{\mathbf{M}}), 0, 1\right). \tag{44}$$

Because $\bar{\mathbf{M}}$ takes intermediate values near transitions between supported and unsupported cells, the term $\bar{\mathbf{M}} \odot (1 - \bar{\mathbf{M}})$ becomes large near observation boundaries. The factor of four rescales its maximum value to one.

The low-confidence boundary weight is

$$\mathbf{B}_{\mathrm{lc}} = \mathbf{M} \odot (1 - \mathbf{C}) \odot \bar{\mathbf{C}}. \tag{45}$$

This term emphasizes low-confidence observed cells that are locally adjacent to more reliable support. The final boundary weight is

$$\mathbf{B} = \mathrm{clip}\left(\mathbf{B}_{\mathrm{tr}} + \mathbf{B}_{\mathrm{lc}}, 0, 1\right). \tag{46}$$

For each latent-grid axis $r \in \{D, H, W\}$, let Δ_r denote the first-order forward difference. Let $\mathbf{B}^{(r)}$ be the average of the boundary weights at the two latent cells adjacent to each difference edge. The boundary-continuity loss is

$$\mathcal{L}_{\mathrm{bd}} = \frac{1}{3} \sum_{r \in \{D,H,W\}} \mathcal{D}\left(\Delta_r \hat{\mathbf{q}}_g, \Delta_r \mathbf{q}_g; \mathbf{B}^{(r)}\right). \tag{47}$$

Unlike total-variation regularization, this loss does not penalize all nonzero gradients. Instead, it matches the predicted latent gradients to the complete target gradients near selected transition regions. It can therefore preserve nonzero geometric structures while reducing artificial boundary discontinuities.

B.5 Complete Objective and Loss Weights

The complete objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{bridge}}\mathcal{L}_{\mathrm{bridge}} + \lambda_{\mathrm{high}}\mathcal{L}_{\mathrm{high}} + \lambda_{\mathrm{low}}\mathcal{L}_{\mathrm{low}}\\ + \lambda_{\mathrm{unobs}}\mathcal{L}_{\mathrm{unobs}} + \lambda_{\mathrm{bd}}\mathcal{L}_{\mathrm{bd}}.\end{aligned} \tag{48}$$

Unless otherwise specified, we use

$$\lambda_{\mathrm{bridge}} = 1.0, \qquad \lambda_{\mathrm{high}} = 0.2, \qquad \lambda_{\mathrm{low}} = 0.5, \tag{49}$$

and

$$\lambda_{\mathrm{unobs}} = 0.5, \qquad \lambda_{\mathrm{bd}} = 0.1. \tag{50}$$

The bridge loss trains the reverse endpoint estimator under samples from the paired conditional bridge. The three regional terms then impose task-specific source preservation, uncertain-region correction, and unsupported-region completion preferences, while the boundary loss improves continuity near local geometric transitions.

C Reliability-Aware Local Geometric Attention: Architecture and Implementation Details

This appendix provides the complete tokenization, neighborhood construction, attention formulation, orientation preprocessing, Transformer block, and implementation details of Reliability-Aware Local Geometric Attention.

C.1 Input Tokenization and Conditioning

The repair network receives the current bridge state $\mathbf{q}_t$, the confidence-aware source endpoint $\tilde{\mathbf{q}}_s$, the binary support map $\mathbf{M}$, the continuous confidence map $\mathbf{C}$, and the bridge timestep t.

Both $\mathbf{q}_t$ and $\tilde{\mathbf{q}}_s$ contain eight channels on a $16 \times 16 \times 16$ latent grid. The support and confidence maps are single-channel fields on the same grid. For each latent-grid location i, we construct the input vector

$$\mathbf{z}_i = [\mathbf{q}_{t,i}; \tilde{\mathbf{q}}_{s,i}; M_i; C_i] \in \mathbb{R}^{18}. \tag{51}$$

The concatenated input is projected to $d_{\text{model}} = 256$ dimensions:

$$\mathbf{x}_i^{(0)} = \mathbf{W}_{\text{in}}\mathbf{z}_i + \phi_p(\mathbf{p}_i) + \phi_t(t/T), \tag{52}$$

where $\mathbf{p}_i \in \mathbb{R}^3$ is the coordinate of token i on the regular latent grid.

The coordinate encoder is

$$\phi_p : \text{Linear}(3, 256) \to \text{SiLU} \to \text{Linear}(256, 256). \tag{53}$$

The discrete timestep is normalized by $T = 1000$ and represented by a 256-dimensional sinusoidal embedding. It is then processed by

$$\phi_t : \text{Linear}(256, 1024) \to \text{SiLU} \to \text{Linear}(1024, 256). \tag{54}$$

The resulting timestep feature is broadcast to all spatial tokens. The complete token sequence contains

$$N_{\text{tok}} = 16^3 = 4096 \tag{55}$$

tokens, each represented by a 256-dimensional feature.

C.2 Fixed Local Grid Neighborhood

Rather than using global attention over all latent tokens, RLGA constructs fixed neighborhoods on the regular 3D latent grid.

The set of candidate offsets is

$$\mathcal{D} = \{(d_z, d_y, d_x) \mid d_z, d_y, d_x \in \{-1, 0, 1\}\}, \tag{56}$$

which contains the 27 offsets of a $3 \times 3 \times 3$ window.

The offsets are deterministically ordered by

$$\left(\|\mathbf{o}\|_2^2, \|\mathbf{o}\|_1, o_z, o_y, o_x\right), \tag{57}$$

namely, first by squared Euclidean distance, then by Manhattan distance, and finally lexicographically. The first $K_{\text{nbr}} = 16$ offsets form the retained set $\mathcal{O}_{16}$.

For a token at $\mathbf{p}_i = (p_{i,z}, p_{i,y}, p_{i,x})$, the local neighborhood is

$$\mathcal{N}(i) = \left\{\text{flatten}\left[\text{clip}\left(\mathbf{p}_i + \mathbf{o}, [0, 15]^3\right)\right] \;\middle|\; \mathbf{o} \in \mathcal{O}_{16}\right\}. \tag{58}$$

The retained neighborhood includes the query token itself, the six axis-aligned neighbors, and the nearest edge-diagonal neighbors. At grid boundaries, coordinates outside the valid range are clamped to the nearest valid coordinate. Repeated neighbor indices can therefore occur for boundary tokens.

The same deterministic neighborhood construction is used in every RLGA block. Unlike point-cloud kNN or radius search, its computational cost does not depend on the spatial density of the source point cloud.

C.3 Complete Reliability-Aware Attention

Let

$$\mathbf{X}^{(l)} = \{\mathbf{x}_i^{(l)}\}_{i=1}^{N_{\text{tok}}} \tag{59}$$

denote the token features at layer l. Each RLGA block uses $H = 8$ attention heads. The per-head dimension is

$$d_h = \frac{d_{\text{model}}}{H} = \frac{256}{8} = 32. \tag{60}$$

Queries, keys, and values are generated using a joint linear projection:

$$[\mathbf{Q}, \mathbf{K}_f, \mathbf{V}] = \mathbf{X}^{(l)}\mathbf{W}_{QKV}. \tag{61}$$

For a neighboring pair (i, j), the geometric descriptor is

$$\mathbf{r}_{ij} = \left[\Delta\mathbf{p}_{ij}, d_{ij}, s_{ij}^n\right] \in \mathbb{R}^5, \tag{62}$$

where

$$\Delta\mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i, \qquad d_{ij} = \|\Delta\mathbf{p}_{ij}\|_2, \qquad s_{ij}^n = \hat{\mathbf{n}}_i^\top \hat{\mathbf{n}}_j. \tag{63}$$

The descriptor is mapped to a head-specific geometric bias:

$$\mathbf{b}_{ij}^{\text{geo}} = \phi_g(\mathbf{r}_{ij}) \in \mathbb{R}^H, \tag{64}$$

where ϕ_g is a two-layer MLP with hidden dimension 64 and output dimension $H = 8$.

When an orientation feature is unavailable for either token, we set

$$s_{ij}^n = 0. \tag{65}$$

Therefore, $\mathbf{r}_{ij}$ remains five-dimensional regardless of orientation availability, and the architecture of ϕ_g remains unchanged.

Reliability first controls which neighboring tokens act as information providers. We define the confidence floor

$$\bar{C}_j = \max(C_j, C_{\min}), \qquad C_{\min} = 0.05. \tag{66}$$

For attention head h, the logit between query token i and neighboring token j is

$$e_{ij}^{(l,h)} = \frac{(\mathbf{q}_i^{(l,h)})^\top \mathbf{k}_j^{(l,h)}}{\sqrt{d_h}} + \lambda_g^{(l)} b_{ij}^{\text{geo},h} + \lambda_c^{(l)} \log(\bar{C}_j), \qquad j \in \mathcal{N}(i). \tag{67}$$

The learnable scales $\lambda_g^{(l)}$ and $\lambda_c^{(l)}$ are maintained independently by each RLGA block and initialized to 1.0 and 0.25, respectively. The geometric bias is head-specific, whereas the two scaling coefficients are shared among heads within the same block.

The normalized attention weights are

$$a_{ij}^{(l,h)} = \frac{\exp(e_{ij}^{(l,h)})}{\sum_{j' \in \mathcal{N}(i)} \exp(e_{ij'}^{(l,h)})}, \qquad j \in \mathcal{N}(i). \tag{68}$$

The head-wise aggregation and output projection are

$$\begin{aligned} \mathbf{o}_i^{(l,h)} &= \sum_{j \in \mathcal{N}(i)} a_{ij}^{(l,h)} \mathbf{v}_j^{(l,h)}, \\ \mathbf{u}_i^{(l)} &= \mathbf{W}_O \,\text{Concat}\left(\mathbf{o}_i^{(l,1)}, \ldots, \mathbf{o}_i^{(l,H)}\right). \end{aligned} \tag{69}$$

Reliability then controls the query token as an information receiver. The query-side update gate is

$$g_i^{\text{upd}} = 1 - C_i, \tag{70}$$

and the residual attention increment is

$$\Delta\mathbf{x}_i^{(l)} = (1 - C_i)\mathbf{u}_i^{(l)}. \tag{71}$$

Therefore, a high-confidence token can provide information to nearby uncertain regions while receiving only a small modification itself. Low-confidence and unsupported tokens receive stronger corrective updates.

Algorithm 1 Reliability-Aware Local Geometric Attention

Require: Token features $\mathbf{X}^{(l)} = \{\mathbf{x}_i^{(l)}\}_{i=1}^{N_{\text{tok}}}$; coordinates $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{N_{\text{tok}}}$; confidences $\mathbf{C} = \{C_i\}_{i=1}^{N_{\text{tok}}}$; optional orientations $\hat{\mathbf{N}} = \{\hat{\mathbf{n}}_i\}_{i=1}^{N_{\text{tok}}}$
Ensure: Residual attention increment $\Delta\mathbf{X}^{(l)}$

$[\mathbf{Q}, \mathbf{K}_f, \mathbf{V}] \leftarrow \mathbf{X}^{(l)}\mathbf{W}_{QKV}$ and reshape into H heads
for each token i **do**
  $\mathcal{N}(i) \leftarrow \mathrm{GridNbr}(\mathbf{p}_i, K_{\text{nbr}})$
  for each $j \in \mathcal{N}(i)$ **do**
    $\Delta\mathbf{p}_{ij} \leftarrow \mathbf{p}_j - \mathbf{p}_i$; $d_{ij} \leftarrow \|\Delta\mathbf{p}_{ij}\|_2$
    $s_{ij}^n \leftarrow \hat{\mathbf{n}}_i^\top \hat{\mathbf{n}}_j$ if available; otherwise $s_{ij}^n \leftarrow 0$
    $\mathbf{r}_{ij} \leftarrow [\Delta\mathbf{p}_{ij}, d_{ij}, s_{ij}^n]$
    $\mathbf{b}_{ij}^{\text{geo}} \leftarrow \phi_g(\mathbf{r}_{ij})$
    $\bar{C}_j \leftarrow \max(C_j, C_{\min})$
    for each attention head h **do**
      Compute $e_{ij}^{(l,h)}$ using equation 11
    end for
  end for
  for each attention head h **do**
    $a_{ij}^{(l,h)} \leftarrow \mathrm{softmax}_{j\in\mathcal{N}(i)}(e_{ij}^{(l,h)})$
    $\mathbf{o}_i^{(l,h)} \leftarrow \sum_{j\in\mathcal{N}(i)} a_{ij}^{(l,h)} \mathbf{v}_j^{(l,h)}$
  end for
  $\mathbf{u}_i^{(l)} \leftarrow \mathbf{W}_O\, \mathrm{Concat}_{h=1}^{H}(\mathbf{o}_i^{(l,h)})$
  $\Delta\mathbf{x}_i^{(l)} \leftarrow (1 - C_i)\mathbf{u}_i^{(l)}$
end for
$\Delta\mathbf{X}^{(l)} \leftarrow \{\Delta\mathbf{x}_i^{(l)}\}_{i=1}^{N_{\text{tok}}}$
return $\Delta\mathbf{X}^{(l)}$

C.4 Complete RLGA Algorithm

Algorithm 1 summarizes the complete RLGA operation.

C.5 Source-Derived Orientation Cues

The orientation cues used in the geometric bias are derived exclusively from the source point set. Ground-truth geometry is not used to construct these features.

Per-point orientation vectors are scattered onto the $64 \times 64 \times 64$ voxel grid. If multiple source points fall into the same voxel, their orientation vectors are averaged and normalized. The orientation grid is then average-pooled to $16 \times 16 \times 16$ and normalized again to align it with the latent tokens.

When explicit source normals are unavailable, we use the centroid-relative orientation proxy

$$\hat{\mathbf{n}}_k = \frac{\mathbf{p}_k - \bar{\mathbf{p}}}{\|\mathbf{p}_k - \bar{\mathbf{p}}\|_2 + \varepsilon_n}, \tag{72}$$

where $\bar{\mathbf{p}}$ is the centroid of the source point set and $\varepsilon_n > 0$ is a small numerical stabilizer. Thus, explicit source normals are not required in our implementation, $s_{ij}^n = 0$ is used only when no valid orientation cue can be constructed. The orientation vectors are not projected into the 256-dimensional token feature space and do not enter the query, key, or value projections. They are used only to compute the scalar orientation consistency term $s_{ij}^n = \hat{\mathbf{n}}_i^\top \hat{\mathbf{n}}_j$.

Roles of $\mathbf{M}$ and $\mathbf{C}$ in RLGA. While both $\mathbf{M}$ and $\mathbf{C}$ are retained in the input token representation, only the continuous confidence $\mathbf{C}$ explicitly modulates the RLGA attention logits and receiver-side update gate. The binary support mask $\mathbf{M}$ remains available through the token representation and the confidence-aware source endpoint, allowing unsupported locations to remain distinguishable from low-confidence observed regions.

D COMPLETE QUANTITATIVE RESULTS

The main paper reports the four metrics most directly related to the geometric repair objective. This section provides the complete eight- metric results, including RGB rendering metrics and standard deviations. Results are reported as mean $\pm$ standard deviation.

D.1 EVALUATION PROTOCOL

Test sets and input protocol. We evaluate object-level reconstruction on a fixed ABO test set and scene-level reconstruction on a fixed 3D-FRONT test set under a single-view protocol. The test sets are fixed before evaluation and shared across all compared methods. For each test instance, all methods use the same input image and test camera configuration and are evaluated using the same metric implementation.

Coordinate normalization. All geometric metrics are computed in the common normalized source coordinate frame used by the reconstruction pipeline. For the source point cloud, we use the center of its axis-aligned bounding box as the origin and isotropically scale the geometry by its maximum bounding-box extent with a padding factor of 1.05. The ground-truth geometry is transformed into the same source-normalized frame using the precomputed geometric transformation from preprocessing. No test-time ICP alignment, rigid pose fitting, or GT-dependent scale fitting is applied to the reconstructed outputs.

Surface sampling and geometric metrics. For each predicted mesh and its corresponding ground-truth mesh, we sample $N_s = 100{,}000$ points from the triangle surfaces using area-weighted surface sampling. Fixed random seeds of 42 and 43 are used for the predicted and ground-truth surfaces, respectively. Let $\mathcal{P}$ and $\mathcal{G}$ denote the resulting predicted and ground-truth point sets. We compute the symmetric Chamfer Distance using Euclidean, rather than squared Euclidean, nearest-neighbor distances:

$$\mathrm{CD}(\mathcal{P},\mathcal{G}) = \frac{1}{2}\left(\frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}}\min_{\mathbf{g}\in\mathcal{G}}\|\mathbf{p}-\mathbf{g}\|_2 + \frac{1}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}}\min_{\mathbf{p}\in\mathcal{P}}\|\mathbf{g}-\mathbf{p}\|_2\right). \tag{73}$$

No additional multiplicative scaling is applied to the reported CD values.

For F-score, precision and recall at distance threshold τ_F are defined as

$$\begin{aligned} P_{\tau_F} &= \frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}}\mathbb{I}\left[\min_{\mathbf{g}\in\mathcal{G}}\|\mathbf{p}-\mathbf{g}\|_2 < \tau_F\right], \\ R_{\tau_F} &= \frac{1}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}}\mathbb{I}\left[\min_{\mathbf{p}\in\mathcal{P}}\|\mathbf{g}-\mathbf{p}\|_2 < \tau_F\right], \end{aligned} \tag{74}$$

and

$$F_{\tau_F} = \frac{2P_{\tau_F}R_{\tau_F}}{P_{\tau_F}+R_{\tau_F}}. \tag{75}$$

We use $\tau_F = 0.01$ in the normalized source coordinate system for all reported F-score results.

Evaluation validity. Metrics are computed independently for each test instance and then aggregated over the fixed test manifest. All evaluated instances produce valid, non-empty geometry for the reported comparisons; therefore, no test sample is excluded because of mesh-extraction or metric-computation failure.

Table 4: **Complete quantitative comparison on ABO.** ↓ indicates lower is better and ↑ indicates higher is better.

Method	Geometry				RGB Rendering			
	CD ↓	F-score (%) ↑	PSNR-N (dB) ↑	LPIPS-N ↓	PSNR (dB) ↑	SSIM (%) ↑	LPIPS ↓	DINO (%) ↓
GaussianAnything (Lan et al., 2025)	0.083 ± 0.0014	51.70 ± 0.92	21.174 ± 0.374	0.201 ± 0.0048	20.332 ± 0.381	90.165 ± 0.472	0.185 ± 0.0049	27.013 ± 0.483
Real3D (Jiang et al., 2025)	0.066 ± 0.0013	57.00 ± 0.85	21.511 ± 0.361	0.179 ± 0.0045	19.741 ± 0.343	91.374 ± 0.485	0.171 ± 0.0047	28.047 ± 0.491
LGM (Tang et al., 2024)	0.074 ± 0.0012	49.20 ± 0.78	20.228 ± 0.326	0.205 ± 0.0049	20.827 ± 0.392	90.621 ± 0.479	0.183 ± 0.0048	23.784 ± 0.457
VoxHammer (Li et al., 2026)	0.047 ± 0.0011	73.10 ± 0.89	20.476 ± 0.339	0.160 ± 0.0042	20.784 ± 0.387	90.650 ± 0.481	0.125 ± 0.0041	15.316 ± 0.394
TRELLIS (Xiang et al., 2025)	0.035 ± 0.0011	84.00 ± 0.95	24.072 ± 0.388	0.106 ± 0.0037	22.237 ± 0.395	92.241 ± 0.493	0.106 ± 0.0036	7.943 ± 0.312
SAM3D (Chen et al., 2026)	0.032 ± 0.0012	84.40 ± 0.97	24.115 ± 0.392	0.102 ± 0.0035	22.736 ± 0.397	92.228 ± 0.491	0.113 ± 0.0038	8.150 ± 0.325
BridgeShape (Kong et al., 2026)	0.028 ± 0.0012	86.00 ± 0.92	24.650 ± 0.380	0.120 ± 0.0037	N/A *	N/A *	N/A *	N/A *
† Points-to-3D (Xia et al., 2026)	0.023 ± 0.0011	89.00 ± 0.99	24.920 ± 0.397	0.086 ± 0.0032	22.864 ± 0.399	92.873 ± 0.497	0.089 ± 0.0033	7.462 ± 0.294
Ours	$\mathbf{0.020 \pm 0.0010}$	$\mathbf{92.00 \pm 0.91}$	$\mathbf{26.890 \pm 0.378}$	$\mathbf{0.076 \pm 0.0031}$	$\mathbf{24.717 \pm 0.381}$	$\mathbf{94.960 \pm 0.498}$	$\mathbf{0.077 \pm 0.0031}$	$\mathbf{6.530 \pm 0.270}$

† Points-to-3D is reimplemented and trained by us following the published method; results may differ from an official implementation.

* BridgeShape is Geometry-only method.

Table 5: **Complete quantitative comparison on 3D-FRONT.** ↓ indicates lower is better and ↑ indicates higher is better.

Method	Geometry				RGB Rendering			
	CD ↓	F-score (%) ↑	PSNR-N (dB) ↑	LPIPS-N ↓	PSNR (dB) ↑	SSIM (%) ↑	LPIPS ↓	DINO (%) ↓
TouchDream (Wang et al., 2026)	0.044 ± 0.0014	48.70 ± 0.71	19.108 ± 0.368	0.263 ± 0.0049	18.536 ± 0.376	84.781 ± 0.469	0.243 ± 0.0048	12.577 ± 0.375
TRELLIS (Xiang et al., 2025)	0.095 ± 0.0015	47.10 ± 0.62	18.512 ± 0.324	0.253 ± 0.0047	18.004 ± 0.318	82.047 ± 0.421	0.234 ± 0.0046	12.114 ± 0.361
SceneGen (Meng et al., 2026)	0.087 ± 0.0013	49.40 ± 0.78	19.429 ± 0.379	0.233 ± 0.0045	18.637 ± 0.382	85.067 ± 0.475	0.235 ± 0.0045	14.737 ± 0.403
MIDI (Huang et al., 2025)	0.076 ± 0.0012	52.10 ± 0.83	21.187 ± 0.386	0.167 ± 0.0036	19.587 ± 0.389	87.213 ± 0.486	0.169 ± 0.0039	14.508 ± 0.398
† Points-to-3D (Xia et al., 2026)	0.039 ± 0.0011	75.50 ± 0.96	21.348 ± 0.394	0.163 ± 0.0033	20.884 ± 0.396	88.152 ± 0.492	0.155 ± 0.0032	9.045 ± 0.287
VoxHammer (Li et al., 2026)	0.052 ± 0.0010	69.70 ± 0.92	20.770 ± 0.381	0.184 ± 0.0038	19.644 ± 0.391	86.315 ± 0.481	0.182 ± 0.0041	18.736 ± 0.486
Ours	$\mathbf{0.035 \pm 0.0010}$	$\mathbf{81.30 \pm 0.98}$	$\mathbf{22.980 \pm 0.397}$	$\mathbf{0.152 \pm 0.0031}$	$\mathbf{22.370 \pm 0.398}$	$\mathbf{89.650 \pm 0.496}$	$\mathbf{0.142 \pm 0.0030}$	$\mathbf{8.210 \pm 0.264}$

† Points-to-3D is reimplemented and trained by us following the published method; results may differ from an official implementation.

Tables 4 and 5 provide the complete eight-metric results and their standard deviations. Because every external method retains its native generation and decoding pipeline, the RGB metrics evaluate end-to-end output quality and are not used to attribute improvements specifically to the bridge repair module.

E Additional Ablation and Region-wise Analysis

E.1 Complete Ablation Results

Table 6 reports the complete eight-metric results for the controlled ablations presented in Sec. 4.3. All variants are trained on the same Toys4K split and evaluated on the same ABO test set, using identical source reconstructions, training configuration, and frozen downstream TRELLIS components.

Table 6: **Complete controlled ablation study on ABO.**

Variant	CD ↓	F-score (%) ↑	PSNR-N (dB) ↑	LPIPS-N ↓	PSNR (dB) ↑	SSIM (%) ↑	LPIPS ↓	DINO (%) ↓
Full Entro-Bridge	**0.020**	**92.0**	**26.89**	**0.076**	**24.717**	**94.96**	0.076	**6.53**
w/o ERSB	0.023	88.5	25.48	0.097	23.46	91.15	0.088	7.68
w/o Observation Mask $\mathbf{M}$	0.021	90.1	26.01	0.085	23.93	92.57	0.079	7.12
w/o Confidence $\mathbf{C}$	0.021	91.2	26.35	0.078	24.25	93.42	0.078	6.75
Variance-Level Endpoint Mixing	0.0207	90.6	26.18	0.084	24.06	93.34	0.079	6.98
w/o Boundary-Continuity Loss	**0.020**	90.8	26.14	0.083	24.11	93.42	0.079	6.96
w/o RLGA	0.026	86.0	24.72	0.118	22.84	89.63	0.105	8.54
w/o RLGA Confidence Bias	**0.020**	90.9	26.31	0.079	24.320	93.72	**0.074**	6.68
w/o RLGA Receiver Gate	0.021	90.2	26.14	0.083	24.150	93.36	0.077	6.91
w/o RLGA Geometry Bias	0.022	89.1	25.67	0.092	23.740	92.41	0.086	7.54
Bridge $\rightarrow$ Noise-Start Conditional Diffusion	0.039	82.3	23.95	0.146	22.05	88.02	0.135	10.85

Table 6 provides a controlled analysis of the two proposed reliability mechanisms and the paired-bridge backbone in Entro-Bridge. First, the ERSB is consistently beneficial. Removing this source-boundary construction increases CD from 0.020 to 0.023, while removing the observation mask $\mathbf{M}$ or continuous confidence $\mathbf{C}$ also degrades reconstruction quality. Replacing the proposed amplitude-level endpoint mixing with variance-level mixing increases CD to 0.0207 and reduces F-score from 92.0% to 90.6%, supporting the proposed parameterization in which confidence jointly controls retained source amplitude and the induced endpoint uncertainty, rather than directly allocating variance alone.

RGRD is supported by complementary loss- and architecture-level evidence. Removing the boundary-continuity loss degrades F-score, normal-rendering metrics, and RGB-rendering metrics while leaving CD unchanged at the reported precision, supporting its role in maintaining local structural consistency near uncertain and observed–unobserved transitions. Replacing RLGA with standard self-attention in the *w/o RLGA* variant causes a substantial degradation, showing that standard attention alone is insufficient to reproduce the full model. Fine-grained RLGA ablations further show that removing these components generally degrades performance, although removing the confidence bias improves RGB LPIPS from 0.076 to 0.074 and leaves CD unchanged at the reported precision. The largest degradation among these three variants occurs when the geometric bias is removed. Finally, replacing the paired bridge backbone with noise-start conditional diffusion produces the largest overall drop, showing the benefit of repairing an informative source endpoint rather than regenerating geometry from Gaussian noise. Overall, the consistent trends across geometry, normal-rendering, and RGB-rendering metrics support the ERSB and reliability-directed learned repair, while the diffusion control validates the choice of paired bridge transport as their backbone.

Confidence robustness. Table 7 evaluates the robustness of Entro-Bridge to confidence perturbations without retraining. We consider uniform, shuffled, and noisy confidence maps, with perturbations applied only to source-supported regions while unsupported locations retain zero confidence.

Table 7: **Robustness to confidence perturbations on ABO.**

Confidence	CD ↓	F-score (%) ↑	Unseen CD ↓	Anchor Drift ↓
Original	0.020	**92.0**	**0.021**	**0.0073**
Uniform	0.021	90.3	0.025	0.0084
Shuffled	0.023	88.8	0.028	0.0096
Noisy ($\sigma = 0.1$)	**0.019**	91.4	0.023	0.0078

As shown in Table 7, mild confidence noise yields mixed changes: CD decreases from 0.020 to 0.019, while F-score decreases from 92.0% to 91.4%, Unseen CD increases from 0.021 to 0.023, and Anchor Drift increases from 0.0073 to 0.0078. Removing spatial confidence variation through uniform confidence further degrades the results, while shuffling the confidence values yields the largest drop (CD 0.023 and Anchor Drift 0.0096). These results indicate that Entro-Bridge is robust to moderate confidence perturbations while benefiting from the spatial correspondence between confidence and source reliability.

Global reconstruction metrics such as Chamfer Distance (CD) and F-score evaluate the final geometry as a whole, yet do not explicitly distinguish between completing previously unobserved geometry and preserving reliable geometry already recovered by the feed-forward reconstructor. We therefore conduct a region-wise analysis on ABO to evaluate this completion–preservation trade-off.

E.2 GT-visible Region Construction

To define observed and unobserved regions independently of our model's internal observation mask, we follow the visible-point construction protocol of Points-to-3D (Xia et al., 2026). Specifically, we uniformly sample a set of surface points $\mathcal{P}_{\mathrm{gt}} = \{\mathbf{p}_i\}_{i=1}^{N}$ from the ground-truth geometry and render its depth map $\mathbf{D}_{\mathrm{gt}}$ from the input camera.

For each GT surface point $\mathbf{p}_i$, we transform it into the input-camera coordinate system and project it onto the image plane. Let z_i denote its projected depth and $\mathbf{u}_i$ its projected image location. We define the visibility indicator as

$$V_i = \begin{cases} 1, & \text{if } |\mathbf{D}_{\mathrm{gt}}(\mathbf{u}_i) - z_i| < \tau_{\mathrm{vis}}, \\ 0, & \text{otherwise}, \end{cases} \tag{76}$$

where $\tau_{\mathrm{vis}} = 0.01$ in the normalized source coordinate system. A GT surface sample is considered visible only when its projection lies within the valid image domain, the rendered GT depth at the projected location is valid, and the depth-consistency criterion in equation 76 is satisfied. Samples with invalid or out-of-image projections are assigned to the unseen subset; geometrically occluded

samples likewise fail the depth-consistency test because the rendered depth corresponds to the closer visible surface. Points satisfying $V_i = 1$ form the GT-visible region,

$$\mathcal{P}_{\mathrm{gt}}^{\mathrm{vis}} = \{\mathbf{p}_i \in \mathcal{P}_{\mathrm{gt}} \mid V_i = 1\}, \tag{77}$$

while the remaining surface points form the GT-unseen region,

$$\mathcal{P}_{\mathrm{gt}}^{\mathrm{unseen}} = \{\mathbf{p}_i \in \mathcal{P}_{\mathrm{gt}} \mid V_i = 0\}. \tag{78}$$

This evaluation partition is constructed solely from the ground-truth geometry and the input camera, and is therefore independent of the confidence map or observation mask used internally by Entro-Bridge.

For region-wise CD and F-score, each reconstructed surface point is assigned the visibility label of its nearest GT surface sample. This produces corresponding predicted subsets $\mathcal{P}_{\mathrm{out}}^{\mathrm{vis}}$ and $\mathcal{P}_{\mathrm{out}}^{\mathrm{unseen}}$ under the same GT-defined partition.

E.3 Region-wise Metrics

Unseen-region completion. We compute CD and F-score between $\mathcal{P}_{\mathrm{out}}^{\mathrm{unseen}}$ and $\mathcal{P}_{\mathrm{gt}}^{\mathrm{unseen}}$. These metrics quantify how accurately each method recovers geometry that is not directly visible from the input view. Lower Unseen CD and higher Unseen F-score indicate better completion of unsupported geometry.

Observed GT Error. To measure the geometric accuracy of the output in the input-supported region, we additionally report a one-sided GT-to-reconstruction error

$$E_{\mathrm{obs}} = \frac{1}{|\mathcal{P}_{\mathrm{gt}}^{\mathrm{vis}}|} \sum_{\mathbf{p}_i \in \mathcal{P}_{\mathrm{gt}}^{\mathrm{vis}}} \min_{\mathbf{q}_j \in \mathcal{P}_{\mathrm{out}}} \|\mathbf{p}_i - \mathbf{q}_j\|_2 . \tag{79}$$

Unlike source-preservation metrics, Observed GT Error evaluates the final reconstruction directly against ground truth. A lower value indicates that the reconstructed geometry remains geometrically accurate in regions supported by the input observation.

Anchor Drift. Observed GT Error alone does not determine whether the reconstruction preserves reliable geometry provided by the feed-forward source. We therefore additionally measure Anchor Drift using a fixed set of high-confidence source-supported anchors. For each test instance, we rank source-supported latent locations satisfying $M_i = 1$ according to their confidence values C_i and select the top 20% as reliable source locations. All source points associated with these selected locations form the anchor set $\mathcal{A}_{\mathrm{anchor}}$. The anchor set is constructed solely from the shared feed-forward source reconstruction and its confidence field, without using the ground-truth geometry or the final reconstruction, and the same anchors are used for all methods for which Anchor Drift is reported.

We define Anchor Drift as

$$D_{\mathrm{anchor}} = \frac{1}{|\mathcal{A}_{\mathrm{anchor}}|} \sum_{\mathbf{p}_i \in \mathcal{A}_{\mathrm{anchor}}} \min_{\mathbf{q}_j \in \mathcal{P}_{\mathrm{out}}} \|\mathbf{p}_i - \mathbf{q}_j\|_2 . \tag{80}$$

A lower Anchor Drift indicates that reliable source-supported geometry undergoes less displacement during completion. Together, Observed GT Error and Anchor Drift distinguish geometric accuracy from simple source retention: an effective method should preserve reliable source geometry without merely retaining inaccurate source predictions.

E.4 Completion–Preservation Results

Table 3 compares Entro-Bridge with the strong point-guided generation baseline Points-to-3D, the vanilla paired-SB baseline, and the variant without the confidence-aware source. The external

comparison evaluates whether Entro-Bridge remains competitive in unseen-region completion and observed-region accuracy, while the controlled bridge variants isolate the effect of explicitly modeling spatially varying source reliability.

The results show that Entro-Bridge achieves a more favorable completion–preservation balance, supporting the effectiveness of explicitly modeling spatially varying source reliability.

F Hyperparameter Sensitivity and Sampling Efficiency

This appendix reports hyperparameter sensitivity and the complete runtime and memory measurements under different numbers of reverse-bridge sampling steps. We analyze the regional regularization scale s_{reg}, the boundary-continuity weight λ_{bd}, and the number of reverse-bridge sampling steps K. These factors respectively control the overall strength of region-aware supervision, the constraint near geometric transition regions, and the reconstruction quality–efficiency trade-off during inference.

In all sensitivity experiments, only one hyperparameter is varied at a time, while the remaining training and evaluation settings are kept unchanged. We report CD and F-score to evaluate overall reconstruction quality. We use Anchor Drift, defined in Appendix E.3, to measure preservation of reliable source-supported geometry.

F.1 Sensitivity Settings

Regional regularization scale. The region-specific loss weights λ_{high}, λ_{low}, and λ_{unobs} jointly control the preservation of reliable geometry, the correction of uncertain geometry, and the completion of unobserved regions. Because these weights are correlated, independently varying all three would make their combined influence difficult to interpret. We therefore introduce an analysis-only scaling factor s_{reg} while preserving their default relative proportions:

$$(\lambda_{\mathrm{high}}, \lambda_{\mathrm{low}}, \lambda_{\mathrm{unobs}}) = s_{\mathrm{reg}}\,(0.2,\, 0.5,\, 0.5)\,. \tag{81}$$

We evaluate

$$s_{\mathrm{reg}} \in \{0.5,\, 1.0,\, 2.0\}, \tag{82}$$

where $s_{\mathrm{reg}} = 1.0$ corresponds to the default loss configuration. Consequently, the unobserved-region completion loss is included in the analysis even though its weight is not varied independently.

Boundary-continuity weight. The boundary-continuity loss supervises local latent variations near observed–unobserved transitions and uncertain geometric boundaries. We evaluate

$$\lambda_{\mathrm{bd}} \in \{0.05,\, 0.10,\, 0.20\}, \tag{83}$$

where $\lambda_{\mathrm{bd}} = 0.10$ is used in the main experiments.

Reverse-bridge sampling steps. The number of reverse-bridge sampling steps controls the number of deterministic updates used to transport the confidence-aware source endpoint toward the estimated complete geometry. We evaluate

$$K \in \{1,\, 3,\, 5,\, 10\}, \tag{84}$$

where $K = 5$ is the default inference configuration.

F.2 Sensitivity Results

RLGA neighborhood size. The neighborhood size K_{nbr} controls the amount of local geometric context aggregated by RLGA. We evaluate

$$K_{\mathrm{nbr}} \in \{7, 16, 27\}, \tag{85}$$

where $K_{\mathrm{nbr}} = 7$ contains the query token and its six axis-aligned neighbors, $K_{\mathrm{nbr}} = 16$ corresponds to the default configuration used in the main experiments, and $K_{\mathrm{nbr}} = 27$ retains the complete $3 \times 3 \times 3$ local neighborhood. All other training and evaluation settings are kept unchanged.

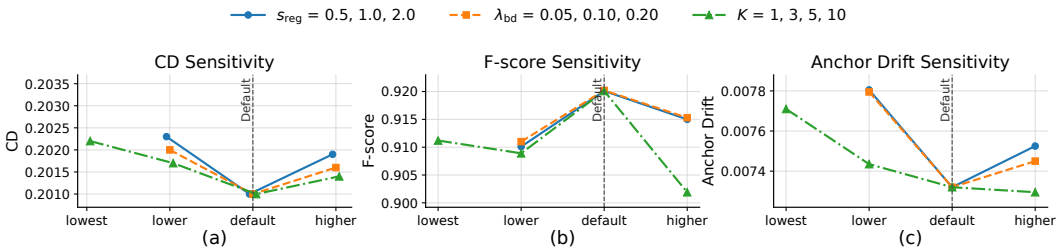

Figure 4: **Hyperparameter sensitivity of Entro-Bridge.** We independently vary the regional regularization scale $s_{\mathrm{reg}} \in \{0.5, 1.0, 2.0\}$, the boundary-continuity weight $\lambda_{\mathrm{bd}} \in \{0.05, 0.10, 0.20\}$, and the number of reverse-bridge sampling steps $K \in \{1, 3, 5, 10\}$. Panels (a), (b), and (c) report CD, F-score, and Anchor Drift, respectively. The horizontal axis represents relative hyperparameter levels rather than a common numerical scale. The lowest category is used only for $K = 1$, whereas lower, default, and higher correspond to the remaining evaluated settings. The dashed vertical line marks the default configuration $(s_{\mathrm{reg}}, \lambda_{\mathrm{bd}}, K) = (1.0, 0.10, 5)$.

F.2.1 Regional regularization scale.

The blue curves in Figs. 4(a)–(c) show the sensitivity to s_{reg}. The default value $s_{\mathrm{reg}} = 1.0$ achieves the lowest CD, the highest F-score, and the lowest Anchor Drift among the evaluated regional scales.

Reducing the scale to 0.5 weakens region-aware supervision, resulting in larger reconstruction error and increased displacement of reliable source geometry. Increasing the scale to 2.0 also degrades all three metrics, indicating that excessive regional regularization can restrict the correction of uncertain geometry and the completion of unobserved regions. These results support $s_{\mathrm{reg}} = 1.0$ as the default setting.

F.2.2 Boundary-continuity weight.

The orange curves in Figs. 4(a)–(c) show the sensitivity to λ_{bd}. The default value $\lambda_{\mathrm{bd}} = 0.10$ achieves the lowest CD, the highest F-score, and the lowest Anchor Drift among the evaluated boundary-loss weights.

Reducing the weight to 0.05 provides weaker guidance around observed–unobserved transitions and uncertain geometric boundaries, leading to lower reconstruction quality and weaker source consistency. Increasing the weight to 0.20 also causes a moderate degradation, suggesting that overly strong boundary regularization restricts the local structural changes required for correction and completion. We therefore use $\lambda_{\mathrm{bd}} = 0.10$ in the main experiments.

F.2.3 Reverse-bridge sampling steps.

The green curves in Figs. 4(a)–(c) show the effect of the number of reverse-bridge sampling steps. Increasing K from 1 to 5 progressively reduces CD and improves F-score. A very short trajectory provides only a coarse approximation of the reverse bridge, whereas $K = 5$ allows the incomplete latent to be refined through several deterministic updates.

Increasing the number of steps further to $K = 10$ produces a marginally lower Anchor Drift while increasing CD and substantially decreasing F-score. This result suggests that additional reverse updates do not necessarily improve the final reconstruction after a sufficiently accurate endpoint estimate has been reached. Although a longer trajectory can make the evolution of reliable source regions more gradual, the small preservation gain at $K = 10$ does not translate into better overall reconstruction quality.

F.2.4 RLGA Neighborhood Size

Table 8 reports the sensitivity of RLGA to the local neighborhood size K_{nbr}.

Table 8: **Sensitivity to the RLGA neighborhood size on ABO.** We vary K_{nbr} while keeping all other training and evaluation settings unchanged. The default configuration is shown in bold.

K_{nbr}	CD ↓	F-score (%) ↑	Anchor Drift ↓
7	0.0203	91.0	0.0080
16 (default)	**0.0201**	**92.0**	**0.0073**
27	0.0199	91.4	0.0077

Performance remains relatively stable across the evaluated neighborhood sizes, with the default $K_{\mathrm{nbr}} = 16$ achieving the highest F-score and lowest Anchor Drift. Reducing the neighborhood to $K_{\mathrm{nbr}} = 7$ provides less local geometric context, resulting in moderately lower reconstruction quality and source preservation. Expanding the neighborhood to the complete $3\times3\times3$ window ($K_{\mathrm{nbr}} = 27$) slightly improves CD from 0.0201 to 0.0199, but reduces F-score and increases Anchor Drift. These results suggest that the default intermediate neighborhood provides sufficient local geometric context while avoiding excessive aggregation from weakly related neighboring locations. Overall, RLGA remains robust to moderate variations in K_{nbr}.

F.3 Step-Wise Sampling Efficiency

Table 9 reports the GPU time and peak memory under different numbers of reverse-bridge sampling steps. These measurements characterize only the reverse-bridge repair process and should be distinguished from the runtime of the complete reconstruction pipeline, which additionally includes condition encoding, frozen TRELLIS Stage-2 generation, and mesh decoding.

All measurements use the same inference configuration as the main experiments. GPU time is measured using CUDA events and averaged over the evaluated samples after warm-up runs.

Table 9: **Efficiency under different numbers of reverse-bridge sampling steps.** The measurements cover reverse-bridge sampling only. Bold indicates the default setting.

Sampling steps K	GPU time (s) ↓	Peak memory (GB) ↓
1	0.064	0.689
3	0.189	0.689
5	**0.314**	**0.689**
10	0.603	0.689

GPU time increases approximately linearly with K, from $0.064\,\mathrm{s}$ at $K = 1$ to $0.603\,\mathrm{s}$ at $K = 10$. In contrast, peak GPU memory remains nearly constant at $0.689\,\mathrm{GB}$ because increasing K repeatedly applies the same bridge repair network without enlarging the intermediate feature representation.

Although $K = 10$ provides a marginal reduction in Anchor Drift, $K = 5$ achieves the best CD and F-score with a bridge-sampling time of only $0.314\,\mathrm{s}$ per sample. We therefore adopt $K = 5$ as the default, providing a favorable balance among reconstruction accuracy, reliable-source preservation, and inference efficiency. The resource consumption of the complete reconstruction pipeline is reported separately in Appendix G.

Overall, the sensitivity and efficiency results support the default configuration $s_{\mathrm{reg}} = 1.0$, $\lambda_{\mathrm{bd}} = 0.10$, and $K = 5$.

G End-to-End Resource and Efficiency Details

We report the computational characteristics of representative 3D completion and generation methods in Table 10. Because the compared methods differ in task definition, representation, resolution, hardware, and timing scope, cross-paper values are provided as contextual references rather than a strict hardware-level ranking.

For Entro-Bridge, all measurements are conducted on the ABO test split using one NVIDIA GeForce RTX 4090 D with batch size 1 and are averaged over 48 samples after five warm-up runs. GPU time is measured using CUDA events. Complete-pipeline timing includes condition encoding, five-step bridge repair, 25-step frozen TRELLIS Stage-2 sampling, and mesh decoding, while excluding data loading, result serialization, visualization rendering, and disk I/O. Peak memory denotes the maximum allocated CUDA memory during inference.

The bridge repair network contains 6.958M trainable parameters, accounting for approximately 0.65% of the complete pipeline. The complete repair-stage measurement is 0.606 s per sample. This value uses a broader scope than the 0.314 s sampler-only result in Table 9, which isolates the iterative reverse-bridge updates. The complete end-to-end pipeline requires 8.361 s per sample and peaks at 10.747 GiB on the stated device.

Table 10: **Reported computational characteristics of representative 3D completion and generation methods.** N/R denotes not reported and N/A denotes not applicable. Cross-paper values are shown for reference because task definitions, hardware, and timing scopes differ.

Method	Task / Representation	Structural Repair Steps ↓	Model Size	Total GFLOPs ↓	Time (s) ↓	Peak Memory	Device / Scope
DiffComplete (Chu et al., 2023)	Voxel completion	100	N/R	15,950	3.20	N/R	32^3, batch size 1
BridgeShape (Kong et al., 2026)	Latent voxel completion	3	N/R	100.8	0.04	N/R	32^3, batch size 1
TRELLIS (Xiang et al., 2025)	3D asset generation	N/A	N/R	N/R	$\sim$10	N/R	Complete asset generation
Points-to-3D (Xia et al., 2026)	Point-guided generation	$25+25$	N/R	N/R	N/R	N/R	Inpainting and refinement
Ours	† Point-guided completion	5	2718.40 MiB	24,448.582*	8.361	10.747 GiB	RTX 4090 D, ABO

†Points-to-3D is reimplemented and trained by us following the published method; results may differ from an official implementation.
*FLOPs are measured using `torch.profiler` with `with_flops=True`. Sparse convolutions, custom CUDA kernels, FlashAttention, and dynamic sparse operations may not be fully covered.

Entro-Bridge performs sparse structural repair in five steps, which is shorter than conventional long-trajectory diffusion completion. The reported BridgeShape runtime is obtained under its own low-resolution voxel setting and is not directly comparable to the complete Entro-Bridge asset-generation pipeline. Likewise, the TRELLIS entry covers native asset generation rather than the proposed repair task.

Points-to-3D does not report wall-clock time, FLOPs, or peak inference memory and has no available official implementation in the supplied experimental record. We therefore do not claim an unverified wall-clock speedup over that method.