STRUCTURE-AWARE PROGRESSIVE REFINEMENT FOR LOCALIZED TEXT-GUIDED POINT CLOUD EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Localized text-guided point-cloud editing aims to modify a target part according to a natural-language instruction while preserving the geometry of the remaining shape. Existing diffusion-based editors typically rely on text conditions and binary part masks to determine what to edit and where to edit. However, such masks do not explicitly encode the geometric relations that should be maintained within structured parts, often leading to asymmetric components, boundary discontinuities, and multi-scale geometric drift during iterative denoising. We propose STELLAR, a structure-aware progressive refinement framework for localized text-guided point-cloud editing. STELLAR progressively refines the editable region by jointly modeling geometric correspondences among parts, the spatial reliability of point-wise corrections, and multi-scale geometric consistency. The framework comprises three coupled components: part-conditioned symmetry projection, which establishes set-wise geometric correspondences under reflection or rotation priors; dynamic spatial reliability calibration, which uses a source-derived spatial reliability field to attenuate symmetry corrections near fragile attachment boundaries; and hierarchical geometric consistency refinement, which regulates source-relative geometric deviations across scales through residual aggregation and bounded cluster-level corrections. We evaluate STELLAR on localized point-cloud editing across multiple object categories using semantic alignment, structural consistency, shape preservation, and multi-scale geometric stability metrics. Our framework provides a unified mechanism for improving the structural reliability of text-guided local point-cloud edits. Project page: `https://stellar-research-site.github.io/`.

1 INTRODUCTION

Localized text-guided point-cloud editing aims to modify a target part while preserving the surrounding geometry. Prior work includes language-driven shape editing with ChangeIt3D (Achlioptas et al., 2023), structural priors for 3D diffusion in SPICE-E (Sella et al., 2024), and localized point-cloud editing with BlendedPC (Sella et al., 2025). BlendedPC combines editing and reconstruction trajectories through mask-based coordinate blending. However, controlling where coordinates change does not explicitly constrain the geometric relations among those coordinates.

This distinction matters for structured objects with repeated legs, bilateral supports, or rotationally arranged components. Partial symmetry provides a useful description of such geometric relations (Bokeloh et al., 2010). Local edits can disrupt these relations even when the protected region is retained. Uniform structural correction is also insufficient: points near attachment boundaries may require weaker intervention, and geometric deviations can occur at different spatial scales. Structural guidance must therefore account for the relation being encouraged, the location of the correction, and its geometric scale.

We propose STELLAR, a structure-aware progressive refinement framework that coordinates these three aspects within reverse diffusion. Part-conditioned symmetry projection constructs geometric targets through set-wise correspondences under category–part priors. Dynamic spatial reliability calibration weights symmetry residuals using a source-derived spatial field, reducing correction strength near protected geometry. Hierarchical geometric consistency refinement aggregates source-relative residuals across fixed spatial partitions and bounds the resulting correction. The refined

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.

clean estimate guides numerical sampling, while reconstruction-based coordinate blending supplies protected coordinates.

Our contributions are:

- We introduce a progressive refinement procedure that coordinates part-conditioned symmetry targets, spatially varying correction strengths, and multi-scale geometric guidance during localized point-cloud editing.
- We combine source-derived spatial reliability with bounded hierarchical residual correction to regulate structural changes while retaining the protected-coordinate reconstruction path.
- We evaluate the components and their combination on chairs, tables, and lamps. The reported results improve structural metrics relative to BlendedPC, with modest runtime overhead and a trade-off in global geometric distortion.

The framework targets edits compatible with the selected structural priors; it does not assume that symmetry or source-shape preservation is appropriate for every instruction.

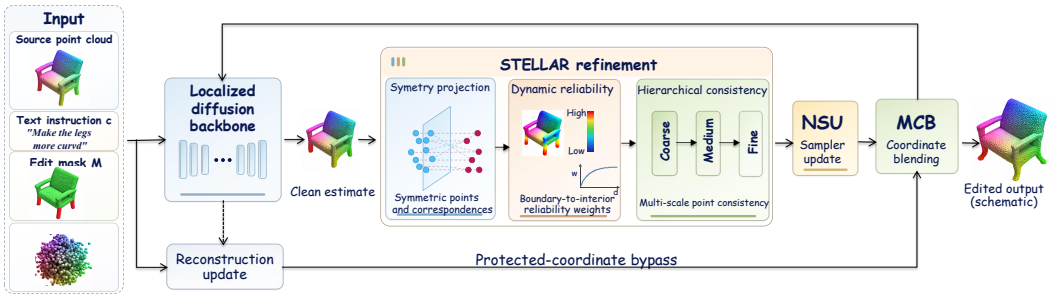

Figure 1: Overview of STELLAR. The inputs are a source point cloud, a text instruction c, and an edit mask M, where $M = 1$ marks editable coordinates and $M = 0$ marks protected coordinates. The localized diffusion backbone has editing and reconstruction branches. STELLAR refines the editable clean estimate through part-conditioned symmetry projection, dynamic spatial reliability calibration, and hierarchical geometric consistency refinement at coarse, medium, and fine scales. NSU denotes numerical sampler update; MCB denotes mask-based coordinate blending. The reconstruction branch bypasses structural refinement and supplies protected coordinates during blending. The chair renderings and edited output are schematic.

2 Related Work

2.1 Localized Text-guided 3D Shape Editing

Text-guided 3D editing aims to modify an existing object according to a language instruction while retaining its identity and the geometry outside the intended editing region. Early methods commonly optimize implicit representations or deformable shape spaces under text guidance, including DualSDF Hao et al. (2020), TextDeformer Gao et al. (2023), and SPAGHETTI Hertz et al. (2022). These methods offer flexible semantic manipulation, yet their optimization-based pipelines can be expensive and often provide limited control over the boundary between modified and preserved geometry.

Recent works have increasingly investigated localized and language-conditioned editing. ShapeTalk provides a dataset of language-described shape differences and introduces the ChangeIt3D framework for language-guided shape editing (Achlioptas et al., 2023). Vox-E (Sella et al., 2023) and SPICE-E (Sella et al., 2024) explore diffusion-based text-guided 3D editing, whereas BlendedPC (Sella et al., 2025) formulates localized point-cloud editing as conditional inpainting and preserves the non-edited context through reconstruction–editing coordinate blending. Preserving the

unedited region does not ensure structural consistency within the edited part. STELLAR complements mask-based localization with part-conditioned symmetry targets, spatially calibrated corrections, and multi-scale residual guidance. These mechanisms refine intermediate clean estimates to address geometric relations within the editable region.

2.2 Structure-aware Generation and Symmetry Priors

Point-cloud generation has progressed from flow-based models Yang et al. (2019) to diffusion-based formulations Luo & Hu (2021); Zeng et al. (2022), substantially improving shape fidelity and diversity. Nevertheless, modeling an object as a single unordered point set can obscure semantic parts and long-range structural relations. Structure-aware shape processing has emphasized that plausible 3D objects are governed not only by local surface geometry but also by compositional regularities, repeated patterns, and spatial relations among parts Mitra et al. (2014).

Several recent generative methods incorporate such structure explicitly. SALAD Koo et al. (2023) models part-level latent diffusion, while DiffFacto Nakayama et al. (2023) supports controllable part-based generation. Quartet of Diffusions Zhou et al. (2026) coordinates global shape, symmetry, semantic parts, and spatial assembly. These works demonstrate the value of structural priors for coherent 3D generation, but their primary objective is compositional generation rather than preserving a language-driven local edit within an existing point cloud.

Symmetry is particularly useful for man-made objects with repeated legs, arms, supports, or counterpart structures. Classical studies on partial and approximate symmetry Bokeloh et al. (2010); Mitra et al. (2006) show that structurally related regions can be recovered without requiring a globally identical object. STELLAR uses a specified category–part symmetry prior to guide the editable subset rather than discovering symmetry transformations from each input instance. Its nearest-neighbor matching updates point correspondences under the selected transformations, not the transformations themselves. The resulting correction is applied through a weighted residual update rather than a hard global symmetry constraint. Instance-level symmetry estimation is therefore distinct from the refinement mechanism considered here.

2.3 Hierarchical Point-cloud Modeling and Multi-scale Geometric Refinement

Point clouds exhibit meaningful organization across resolutions: coarse scales capture global silhouette and part layout, while finer scales describe local surface details and attachment geometry. PointNet++ (Qi et al., 2017) learns hierarchical features through nested spatial neighborhoods. Point Transformer (Zhao et al., 2021) uses attention to aggregate geometric context among points, whereas PointMamba (Liang et al., 2024) uses space-filling curves and a non-hierarchical state-space encoder for efficient point-cloud analysis. These approaches provide complementary mechanisms for geometric representation learning.

PointNSP (Meng et al., 2026) develops next-scale level-of-detail prediction: low-resolution point sets establish global shape structure, and finer scales recover detailed geometry. Invaria (Chang et al., 2026) also studies scale and density invariance in point-cloud representations. In contrast, STELLAR uses a source-based hierarchy to regularize localized editing during reverse diffusion. It fixes the spatial partitions before sampling, computes cluster-level residuals from the current editable proposal, and averages these residuals across hierarchy levels before applying a bounded correction.This procedure provides multi-scale structural guidance by aggregating source-relative residuals, with norm clipping controlling the magnitude of the hierarchical correction before guidance weighting.

3 Method

3.1 Problem Setup and Localized Editing

Let $\mathbf{X}^{\mathrm{src}} = \{\mathbf{x}_i^{\mathrm{src}}\}_{i=1}^{N} \in \mathbb{R}^{N\times 3}$ be a source point cloud with N points, where $\mathbf{x}_i^{\mathrm{src}} \in \mathbb{R}^3$ is the coordinate of point i. Given a text instruction c and a binary part mask $\mathbf{M} = \{m_i\}_{i=1}^{N}$, where $m_i = 1$ marks an editable point and $m_i = 0$ marks a protected point, the task is to produce an

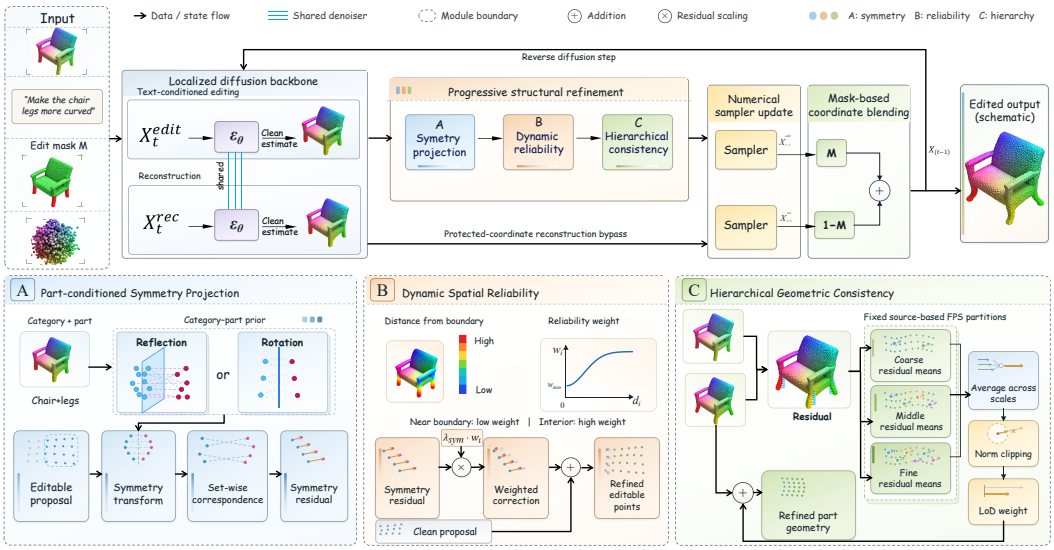

Figure 2: STELLAR workflow at reverse-diffusion step t. The inputs are a source point cloud, a text instruction c, and a binary edit mask $\mathbf{M}$. The editing and reconstruction trajectories share a noise predictor; D_θ in the diagram denotes $\boldsymbol{\epsilon}_\theta$ in the text. The editable clean estimate undergoes (A) part-conditioned symmetry projection, (B) spatial reliability calibration, and (C) hierarchical geometric consistency refinement. FPS denotes farthest point sampling, and LoD denotes level of detail. The reconstruction trajectory bypasses these three modules and supplies protected coordinates during mask-based blending. Chair renderings and overlays are schematic, not experimental outputs.

edited cloud $\mathbf{X}^{\text{out}}$ that follows c while preserving the protected geometry. We denote the editable and protected index sets by $\mathcal{I}_e$ and $\mathcal{I}_p$, respectively, and their source subsets by $\mathbf{X}_e^{\text{ref}}$ and $\mathbf{X}_p^{\text{ref}}$:

$$\begin{cases} \mathcal{I}_e = \{i : m_i = 1\}, & \mathbf{X}_e^{\text{ref}} = \{\mathbf{x}_i^{\text{src}} : i \in \mathcal{I}_e\}, \\ \mathcal{I}_p = \{i : m_i = 0\}, & \mathbf{X}_p^{\text{ref}} = \{\mathbf{x}_i^{\text{src}} : i \in \mathcal{I}_p\}. \end{cases} \tag{1}$$

We use the localized diffusion setting of BlendedPC (Sella et al., 2025). Let $\mathbf{X}_0$ be a clean point cloud, $\mathbf{X}_t$ its noisy version at diffusion step $t \in \{1, \ldots, T\}$, and T the total number of steps. The coefficient $\bar{\alpha}_t \in (0, 1]$ is the cumulative fraction of retained clean signal at step t. Gaussian noise $\boldsymbol{\epsilon}$ has zero mean and identity covariance $\mathbf{I}$. The forward process is

$$\mathbf{X}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{X}_0 + \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

At inference, the masked source context is $\mathbf{X}^M = \mathbf{X}^{\text{src}} \odot (1 - \mathbf{M})$, where $\odot$ is element-wise multiplication and the mask is broadcast over the three coordinate dimensions. The noise predictor $\boldsymbol{\epsilon}_\theta$, with parameters θ, receives the noisy cloud, step, text, and masked context. As in BlendedPC, an editing trajectory $\mathbf{X}_t^{\text{edit}}$ and a reconstruction trajectory $\mathbf{X}_t^{\text{rec}}$ start from shared noise and use the same predictor. The editing trajectory uses c; the reconstruction trajectory uses the source shape and an empty text condition.

After the numerical sampler advances both trajectories from t to $t-1$, their outputs $\mathbf{X}_{t-1}^{\text{edit}}$ and $\mathbf{X}_{t-1}^{\text{rec}}$ are combined as

$$\mathbf{X}_{t-1} = \mathbf{M} \odot \mathbf{X}_{t-1}^{\text{edit}} + (1 - \mathbf{M}) \odot \mathbf{X}_{t-1}^{\text{rec}}. \tag{3}$$

Thus, the mask determines which trajectory supplies each coordinate, but does not specify geometric relations among editable points.

3.2 STRUCTURE-AWARE PROGRESSIVE REFINEMENT

At reverse step t, STELLAR first forms the editing branch's clean estimate $\widetilde{\mathbf{X}}_0^{(t)}$:

$$\widetilde{\mathbf{X}}_0^{(t)} = \frac{\mathbf{X}_t^{\text{edit}} - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_\theta(\mathbf{X}_t^{\text{edit}}, t, c, \mathbf{X}^M)}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

Here, $\widetilde{\mathbf{x}}_i^{(t)}$ denotes point i of this estimate. The following three modules refine its editable points in order: a symmetry target specifies a geometric relation, a reliability field controls the local strength of the corresponding correction, and a hierarchy constrains source-relative deviations across scales.

Part-conditioned symmetry projection. For a selected category–part pair p, let $\mathcal{G}_p = \{g_1, \ldots, g_{K_g}\}$ be a nonempty set of K_g prescribed symmetry transformations in the source cloud's canonical coordinate frame. A reflection is defined by a plane $\mathbf{n}^\top \mathbf{x} + b = 0$, where $\mathbf{n}$ is its unit normal, b its offset, and $\mathbf{x} \in \mathbb{R}^3$ a point. Alternatively, a rotational prior uses a unit axis $\mathbf{a}$ through the canonical origin and an integer rotational order $K_r \geq 2$. The two transformation forms are

$$\begin{cases} g_{\mathbf{n},b}(\mathbf{x}) = \mathbf{x} - 2(\mathbf{n}^\top \mathbf{x} + b)\mathbf{n}, & \|\mathbf{n}\|_2 = 1, \\ g_k(\mathbf{x}) = \mathbf{R}\left(\mathbf{a}, \dfrac{2\pi k}{K_r}\right)\mathbf{x}, & k = 1, \ldots, K_r - 1. \end{cases} \tag{5}$$

Here, $\mathbf{R}(\mathbf{a}, \phi)$ is the three-dimensional rotation matrix for angle ϕ around $\mathbf{a}$. The chosen prior must be compatible with the object's alignment, structure, and requested edit.

For each transformation $g \in \mathcal{G}_p$ and editable point $i \in \mathcal{I}_e$, let $\pi_g(i)$ be the index of the nearest editable point to the transformed current proposal. The symmetry target $\mathbf{q}_i^{\text{sym},(t)}$ averages the current point with these correspondences mapped back by the inverse transformation g^{-1}:

$$\begin{aligned} \pi_g(i) &= \arg\min_{j \in \mathcal{I}_e} \left\| g(\widetilde{\mathbf{x}}_i^{(t)}) - \widetilde{\mathbf{x}}_j^{(t)} \right\|_2^2, \\ \mathbf{q}_i^{\text{sym},(t)} &= \frac{1}{2}\left[\widetilde{\mathbf{x}}_i^{(t)} + \frac{1}{K_g} \sum_{g \in \mathcal{G}_p} g^{-1}\left(\widetilde{\mathbf{x}}_{\pi_g(i)}^{(t)} \right) \right]. \end{aligned} \tag{6}$$

These are set-wise nearest-neighbor assignments, not fixed point pairs or a bijection. They are recomputed at every reverse step and may change between steps. The target guides a residual correction; it does not replace the point directly.

Dynamic spatial reliability calibration. Assume both $\mathcal{I}_e$ and $\mathcal{I}_p$ are nonempty. For editable point i, let d_i be its source-cloud distance to the nearest protected point. We normalize it by s_{box}, the diagonal length of the source cloud's axis-aligned bounding box, using a small $\varepsilon > 0$ for numerical stability:

$$\begin{cases} d_i = \min\limits_{j \in \mathcal{I}_p} \|\mathbf{x}_i^{\text{src}} - \mathbf{x}_j^{\text{src}}\|_2, \\ \widetilde{d}_i = \dfrac{d_i}{s_{\text{box}} + \varepsilon}. \end{cases} \tag{7}$$

Let w_i be the reliability weight at point i. Its lower bound $w_{\min} \in [0, 1]$, transition width $\sigma > 0$, and curve exponent $\gamma > 0$ define

$$w_i = w_{\min} + (1 - w_{\min})\left[1 - \exp\left(-\frac{\widetilde{d}_i^2}{2\sigma^2} \right) \right]^\gamma. \tag{8}$$

The symmetry correction strength $\lambda_{\text{sym}} \geq 0$ then produces the reliability-calibrated proposal $\overline{\mathbf{x}}_i^{(t)}$:

$$\overline{\mathbf{x}}_i^{(t)} = \widetilde{\mathbf{x}}_i^{(t)} + \lambda_{\text{sym}} w_i \left(\mathbf{q}_i^{\text{sym},(t)} - \widetilde{\mathbf{x}}_i^{(t)} \right), \qquad i \in \mathcal{I}_e. \tag{9}$$

The source-derived distances and weights remain fixed during sampling, whereas the symmetry target and its residual change with the current proposal. Near protected geometry, smaller w_i reduces the intervention. Distance is only a proxy for attachment sensitivity; it does not guarantee boundary continuity.

Hierarchical geometric consistency refinement. The symmetry target alone does not control shared geometric drift across spatial scales. We therefore build $L \geq 1$ fixed source-based levels. At level $l \in \{1, \ldots, L\}$, s_l is the number of anchors, with $1 \leq s_l \leq |\mathcal{I}_e|$. Farthest point sampling, denoted FPS, selects the anchor set $\mathbf{A}^{(l)} = \{\mathbf{a}_k^{(l)}\}_{k=1}^{s_l}$. The function $c_l(i)$ assigns editable point i to the index k of its nearest source-based anchor:

$$\begin{cases} \mathbf{A}^{(l)} = \mathrm{FPS}(\mathbf{X}_e^{\mathrm{ref}}, s_l), \\ c_l(i) = \arg\min\limits_{1 \leq k \leq s_l} \|\mathbf{x}_i^{\mathrm{src}} - \mathbf{a}_k^{(l)}\|_2^2. \end{cases} \tag{10}$$

The anchor sets and assignments remain fixed during reverse diffusion. This construction requires the source and current editable clouds to retain consistent point indices; it does not perform semantic rematching after reordering or resampling.

At step t, the source-relative residual $\mathbf{r}_i^{(t)}$ compares source point i with its reliability-calibrated proposal. For cluster k at level l, $\boldsymbol{\mu}_{l,k}^{(t)}$ is the mean of those residuals. The indicator $\mathbf{1}[\cdot]$ equals one when its condition is true and zero otherwise:

$$\begin{aligned} \mathbf{r}_i^{(t)} &= \mathbf{x}_i^{\mathrm{src}} - \overline{\mathbf{x}}_i^{(t)}, \\ \boldsymbol{\mu}_{l,k}^{(t)} &= \frac{\sum_{i \in \mathcal{I}_e} \mathbf{1}[c_l(i) = k] \mathbf{r}_i^{(t)}}{\sum_{i \in \mathcal{I}_e} \mathbf{1}[c_l(i) = k] + \varepsilon}. \end{aligned} \tag{11}$$

For each editable point, the cross-level residual $\boldsymbol{\Delta}_i^{\mathrm{hgc},(t)}$ averages the means of its assigned clusters. Let $\mathcal{C}_{\delta_{\max}}$ clip a vector's Euclidean norm using a maximum displacement $\delta_{\max} > 0$, and let $\lambda_{\mathrm{lod}} \geq 0$ be the level-of-detail correction strength. The final editable point $\widehat{\mathbf{x}}_i^{(t)}$ is

$$\begin{aligned} \boldsymbol{\Delta}_i^{\mathrm{hgc},(t)} &= \frac{1}{L} \sum_{l=1}^{L} \boldsymbol{\mu}_{l,c_l(i)}^{(t)}, \\ \mathcal{C}_{\delta_{\max}}(\mathbf{v}) &= \mathbf{v} \min\left(1, \frac{\delta_{\max}}{\|\mathbf{v}\|_2 + \varepsilon}\right), \\ \widehat{\mathbf{x}}_i^{(t)} &= \overline{\mathbf{x}}_i^{(t)} + \lambda_{\mathrm{lod}} \, \mathcal{C}_{\delta_{\max}}\left(\boldsymbol{\Delta}_i^{\mathrm{hgc},(t)}\right), \quad i \in \mathcal{I}_e. \end{aligned} \tag{12}$$

Here, $\mathbf{v} \in \mathbb{R}^3$ is an arbitrary correction vector. Clipping bounds each cross-level correction before multiplication by λ_{lod}; it does not bound the cumulative displacement over all reverse steps. Because the residual is source-relative, this module may oppose large edits that the instruction genuinely requires.

The refined editable points form the editable portion of $\widehat{\mathbf{X}}_0^{(t)}$; its protected points remain those of $\widetilde{\mathbf{X}}_0^{(t)}$. The numerical sampler uses this full refined clean estimate to advance the editing trajectory. The reconstruction trajectory bypasses all three modules, and the two sampled trajectories are blended by Eq. equation 3 at each reverse step.

4 Experiments

4.1 Experimental Settings and Baselines

We evaluate STELLAR on the ShapeTalk benchmark Achlioptas et al. (2023), which is derived from ShapeNet Chang et al. (2015). We focus on three structurally diverse domains: chair, table, and lamp. To accommodate their different geometric distributions, we adopt a category-aware expert protocol and train a localized text-conditioned diffusion editor for each category.

Following the protocol in the original study, the diffusion editors are trained for 500 epochs with a learning rate of 1×10^{-4} and batch size 32. The best checkpoint is selected using held-out validation performance, while the test split remains unseen until final reporting.

We compare STELLAR with ChangeIt3D Achlioptas et al. (2023), SPICE-E Sella et al. (2024), and BlendedPC Sella et al. (2025). For a strict comparison with BlendedPC, we fix the checkpoint, initial

Table 1: Quantitative evaluation on ShapeTalk. Higher values are better for CLIP Sim, while lower values are better for all remaining metrics.

Method	CLIP Sim ↑	CLIP Dir ↓	SDI-CD ↓	LoD-CD ↓	GD ↓	l-GD ↓	FPD ↓	Class-D ↓
ChangeIt3D Achlioptas et al. (2023)	0.2100	1.020	0.7643	0.2791	0.650	0.190	183.02	0.1800
SPICE-E Sella et al. (2024)	0.2500	1.010	0.7842	0.2769	1.840	0.310	390.02	0.2400
BlendedPC Sella et al. (2025)	0.2600	0.990	0.6742	0.2688	**0.340**	**0.070**	33.64	0.0500
STELLAR	**0.2607**	**0.987**	**0.6422**	**0.2591**	0.348	**0.070**	**33.17**	**0.0497**

noise, source point cloud, edit mask, text instruction, and sampling schedule. The only difference is whether STELLAR's structural refinement is activated. Before metric computation, all outputs are canonically aligned and resampled to the same point cardinality.

4.2 Evaluation Metrics

We evaluate semantic fidelity, symmetry, identity preservation, and multi-scale geometric stability. Higher CLIP Similarity and lower values of the other metrics are preferred.

CLIP Similarity (CLIP Sim) measures the alignment between multi-view renderings of the edited output and the target instruction using CLIP features (Radford et al., 2021; Sella et al., 2025).

CLIP Directional Distance (CLIP Dir) measures the disagreement between the source-to-edit direction in CLIP image-feature space and the corresponding direction in CLIP text-feature space, following Sella et al. (2025).

SDI-CD measures self-symmetry consistency using the Chamfer Distance between the edited point set and its symmetry-transformed counterpart. This is a variant of the symmetry discrepancy index introduced by Zhou et al. (2026); the original metric instead compares a shape with its reconstruction from a fundamental domain.

LoD-CD is our multi-resolution geometric discrepancy metric. It averages the Chamfer Distance between the output and the evaluation reference at resolutions of 256, 512, and 1024 points.

Geometric Distance (GD) measures the Chamfer Distance between the source and edited point clouds. **Localized Geometric Distance (l-GD)** evaluates geometric change outside the editable mask. These metrics follow ShapeTalk and the localized evaluation protocol of BlendedPC (Achlioptas et al., 2023; Sella et al., 2025).

Fréchet Point Cloud Distance (FPD) measures distributional discrepancy in point-cloud feature space. It was introduced by Shu et al. (2019) and subsequently used for shape-editing evaluation (Achlioptas et al., 2023; Sella et al., 2025).

Class Distortion (Class-D) measures the change in a shape classifier's probability assigned to the source category between the source and edited shapes (Achlioptas et al., 2023; Sella et al., 2025).

4.3 Quantitative Results

Compared with BlendedPC, STELLAR reduces SDI-CD from 0.6742 to 0.6422 and LoD-CD from 0.2688 to 0.2591 (Table 1). These structural improvements accompany smaller changes in the semantic metrics: CLIP Sim increases from 0.2600 to 0.2607, and CLIP Dir decreases from 0.990 to 0.987. FPD decreases from 33.64 to 33.17, while Class-D decreases from 0.0500 to 0.0497. Together, these measurements support the intended role of STELLAR as a structural refinement of the localized editing trajectory.

The improvements do not extend to every measure of source preservation. GD increases from 0.340 to 0.348, while l-GD remains 0.070 at the reported precision. This indicates a trade-off between the measured structural consistency and global source-shape preservation. Further per-instance analysis would be needed to establish whether the additional displacement is concentrated in intended edits or unwanted geometric changes.

Table 2: Ablation study of the STELLAR inference framework. Sym., Rel., and HGC denote part-conditioned symmetry projection, dynamic spatial reliability calibration, and hierarchical geometric consistency refinement, respectively. Original denotes the BlendedPC backbone without these refinement modules. Under the stated formulation, Rel. is not applicable (n/a) when the symmetry correction is removed.

Setting	Sym.	Rel.	HGC	CLIP Sim ↑	CLIP Dir ↓	SDI-CD ↓	LoD-CD ↓	GD ↓	l-GD ↓	FPD ↓	Class-D ↓
Original	×	×	×	0.2600	0.990	0.6742	0.2688	**0.340**	**0.070**	33.64	0.0500
STELLAR w/o Rel.	✓	×	✓	0.2602	0.989	0.6548	0.2637	0.343	0.071	33.42	0.0499
STELLAR w/o HGC	✓	✓	×	0.2604	0.988	0.6519	0.2631	0.345	0.071	33.35	0.0498
STELLAR w/o Sym.	×	n/a	✓	0.2603	0.989	0.6635	0.2648	0.343	0.071	33.39	0.0499
STELLAR	✓	✓	✓	**0.2607**	**0.987**	**0.6422**	**0.2591**	0.348	**0.070**	**33.17**	**0.0497**

Table 3: Inference efficiency and output validity. Time and peak memory consumption are averaged over valid generated samples. Valid denotes the percentage of generated samples that meet the evaluation's output-validity criterion.

Method	Time (s) ↓	Memory (GB) ↓	Valid (%) ↑
ChangeIt3D	19.362	4.659	99
SPICE-E	**15.421**	**2.376**	92
BlendedPC	17.782	3.429	**100**
STELLAR	18.255	3.431	**100**

4.4 Ablation Study

We examine the contributions of part-conditioned symmetry projection (Sym.), dynamic spatial reliability calibration (Rel.), and hierarchical geometric consistency refinement (HGC) through component ablations. All variants use the same diffusion backbone, input point clouds, editing instructions, random seeds, and sampling steps.

Removing Sym. increases SDI-CD from 0.6422 to 0.6635 (Table 2), consistent with the role of symmetry correspondences in regularizing repeated structures. Removing HGC increases LoD-CD from 0.2591 to 0.2631, supporting the use of multi-scale residual correction. Without Rel., SDI-CD increases to 0.6548 and l-GD increases from 0.070 to 0.071. These changes support spatial calibration of symmetry guidance, although the aggregate metrics do not directly measure attachment continuity.

Rel. weights the correction produced by Sym. and does not introduce an independent displacement when that correction is absent. The variant without Sym. therefore evaluates HGC without an active reliability-weighted symmetry update. The complete configuration gives the lowest SDI-CD and LoD-CD among the listed variants, while retaining the GD trade-off discussed above.

4.5 Qualitative Comparison

Figure 3 compares ChangeIt3D, BlendedPC, and STELLAR on localized chair editing. Under the same input shapes and instruction "legs are more curved", STELLAR produces more coherent part-level modifications while preserving the non-edited geometry.

Detailed comparisons between BlendedPC and STELLAR are shown in the right panel of Figure 3. The red dashed boxes highlight local geometric differences between the two methods.

4.6 Inference Efficiency and Output Validity

Table 3 shows that STELLAR adds 0.473 seconds per sample over BlendedPC (17.782 to 18.255 seconds; approximately 2.7%), with peak memory increasing from 3.429 to 3.431 GB. Both methods achieve 100% output validity under the evaluation criterion, which does not establish semantic accuracy or structural correctness.

Discussion. STELLAR targets localized edits compatible with the selected category–part priors. Its formulation assumes suitable canonical alignment and consistent point indexing. Source-derived

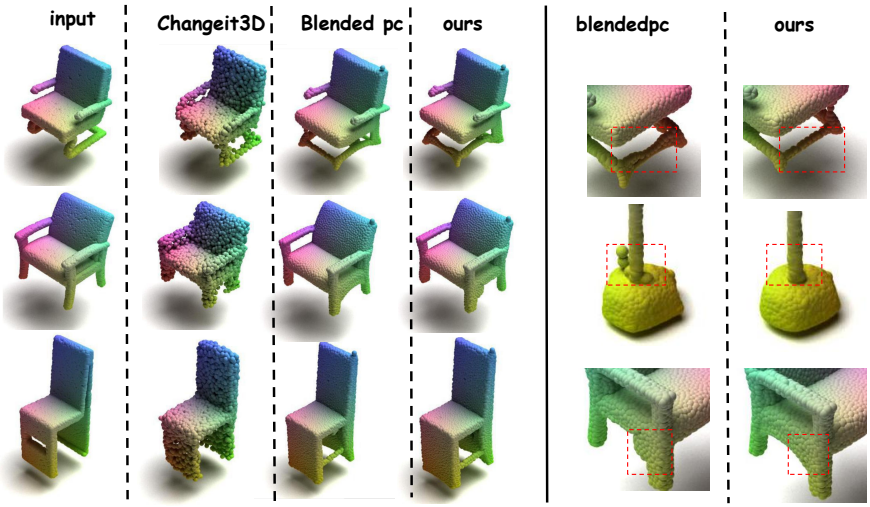

Figure 3: Overall and detailed visual comparison of localized chair editing. Left: full-chair results from ChangeIt3D, BlendedPC, and STELLAR for the same input shapes and editing instruction. Right: close-up comparisons of BlendedPC and STELLAR, highlighting chair-leg geometry and attachment regions. Red dashed boxes mark the local details of interest.

reliability weights remain fixed during sampling, while source-relative HGC corrections can oppose large or asymmetric edits despite bounded update magnitudes. The reported gains therefore apply to the evaluated setting and do not establish robustness to incorrect priors, misalignment, or changing point correspondences.

5 Conclusion

We presented STELLAR, a structure-aware refinement framework for localized text-guided point-cloud editing. Its central idea is to complement an editing mask with structural corrections calibrated by spatial reliability and geometric scale. Part-conditioned symmetry projection, dynamic spatial reliability calibration, and hierarchical geometric consistency refinement operate on editable clean estimates within reverse diffusion. The reported evaluation shows improved structural and multi-scale consistency relative to BlendedPC, with modest runtime overhead and slightly higher global geometric distortion. The framework remains dependent on suitable category–part priors, motivating future work on instruction-aware prior selection.

AI Use Statement

Generative AI tools assisted manuscript drafting and revision, English-language editing, LaTeX organization, and preparation of schematic figures. The authors are responsible for the technical content, experimental results, citations, and final submission.

Reproducibility Statement

The method section and appendix describe the structural refinement operators, training and inference configurations, and evaluation protocol. The appendix provides additional implementation details and experimental settings to support reproducibility.

REFERENCES

Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, and Leonidas Guibas. ShapeTalk: A language dataset and framework for 3D shape edits and deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12685–12694, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Achlioptas_ShapeTalk_A_Language_Dataset_and_Framework_for_3D_Shape_Edits_CVPR_2023_paper.html.

Martin Bokeloh, Michael Wand, and Hans-Peter Seidel. A connection between partial symmetry and inverse procedural modeling. In *ACM SIGGRAPH 2010 Papers*, 2010. URL https://doi.org/10.1145/1833349.1778841. Article 104.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository, 2015. URL https://arxiv.org/abs/1512.03012. arXiv:1512.03012.

Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, Dariu Gavrila, and Holger Caesar. Invaria: Learning scale and density invariance in point clouds via next-resolution prediction, 2026. URL https://arxiv.org/abs/2605.15923. arXiv:2605.15923.

William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. TextDeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. URL https://doi.org/10.1145/3588432.3591552.

Zekun Hao, Hadar Averbuch-Elor, Noah Snavely, and Serge Belongie. DualSDF: Semantic shape manipulation using a two-level representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7631–7641, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Hao_DualSDF_Semantic_Shape_Manipulation_Using_a_Two-Level_Representation_CVPR_2020_paper.html.

Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. SPAGHETTI: Editing implicit shapes through part aware generation. *ACM Transactions on Graphics*, 41(4), 2022. URL https://doi.org/10.1145/3528223.3530084. Article 106.

Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. SALAD: Part-level latent diffusion for 3D shape generation and manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14441–14451, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Koo_SALAD_Part-Level_Latent_Diffusion_for_3D_Shape_Generation_and_Manipulation_ICCV_2023_paper.html.

Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. PointMamba: A simple state space model for point cloud analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/395371f778ebd4854b88521100af30ad-Abstract-Conference.html.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3D point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2837–2845, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Diffusion_Probabilistic_Models_for_3D_Point_Cloud_Generation_CVPR_2021_paper.html.

Ziqiao Meng, Qichao Wang, Zhiyang Dou, Zixing Song, Zhipeng Zhou, Irwin King, and Peilin Zhao. PointNSP: Autoregressive 3D point cloud generation with next-scale level-of-detail prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31450–31461, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Meng_PointNSP_Autoregressive_3D_Point_Cloud_Generation_with_Next-Scale_Level-of-Detail_Prediction_CVPR_2026_paper.html.

Niloy J. Mitra, Leonidas J. Guibas, and Mark Pauly. Partial and approximate symmetry detection for 3D geometry. *ACM Transactions on Graphics*, 25(3):560–568, 2006. URL https://doi.org/10.1145/1141911.1141924.

Niloy J. Mitra, Michael Wand, Hao Zhang, Daniel Cohen-Or, Vladimir Kim, and Qi-Xing Huang. Structure-aware shape processing. In *ACM SIGGRAPH 2014 Courses*, 2014. URL https://doi.org/10.1145/2614028.2615401. Article 13.

George Kiyohiro Nakayama, Mikaela Angelina Uy, Jiahui Huang, Shi-Min Hu, Ke Li, and Leonidas Guibas. DiffFacto: Controllable part-based 3D point cloud generation with cross diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14257–14267, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Nakayama_DiffFacto_Controllable_Part-Based_3D_Point_Cloud_Generation_with_Cross_Diffusion_ICCV_2023_paper.html.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-E: Text-guided voxel editing of 3D objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 430–440, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Sella_Vox-E_Text-Guided_Voxel_Editing_of_3D_Objects_ICCV_2023_paper.html.

Etai Sella, Gal Fiebelman, Noam Atia, and Hadar Averbuch-Elor. Spice-E: Structural priors in 3D diffusion using cross-entity attention. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. URL https://doi.org/10.1145/3641519.3657461.

Etai Sella, Noam Atia, Ron Mokady, and Hadar Averbuch-Elor. Blended point cloud diffusion for localized text-guided shape editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19119–19129, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Sella_Blended_Point_Cloud_Diffusion_for_Localized_Text-guided_Shape_Editing_ICCV_2025_paper.html.

Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3D point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. PointFlow: 3D point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4541–4550, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Yang_PointFlow_3D_Point_Cloud_Generation_With_Continuous_Normalizing_Flows_ICCV_2019_paper.pdf.

Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: Latent point diffusion models for 3D shape generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10021–10039, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/40e56dabe12095a5fc44a6e4c3835948-Abstract.html.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*

(ICCV), pp. 16259–16268, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_Point_Transformer_ICCV_2021_paper.html.

Chenliang Zhou, Fangcheng Zhong, Weihao Xia, Albert Miao, Canberk Baykal, and Cengiz Oztireli. Quartet of diffusions: Structure-aware point cloud generation through part and symmetry guidance. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3e0945e16fa4249e6850ea42390e955f-Abstract-Conference.html.

SUPPLEMENTARY MATERIAL

This appendix reports supplementary quantitative results first, then evaluation settings, mathematical and implementation details, and finally qualitative visualizations and limitations. The operator definitions use the same notation as the main text.

Appendix order.

- Appendix A: ablation, category-level and part-level results, followed by inference efficiency.
- Appendix B: paired evaluation protocol, metrics, priors, model settings, and parameter selection.
- Appendix C: notation, diffusion backbone, symmetry, reliability, hierarchical correction, sampler update, and reproducibility details.
- Appendix D: failure-case analysis and all supplementary qualitative visualizations.

A ADDITIONAL EXPERIMENTAL RESULTS

The following tables supplement the main-text comparisons. The ablation and category-level tables are part of this appendix and are anchored below their respective descriptions.

Table shading. In quantitative comparisons, rose marks the best metric value and peach the second-best within each comparison group; ties receive the same shade. Pale row bands in all tables aid reading and do not imply a ranking.

A.1 ABLATION STUDY

The ablation removes one structural operator at a time while keeping the checkpoint, samples, prompts, masks, random seeds, and sampler settings fixed. The full configuration gives the lowest SDI-CD, LoD-CD, FPD, and Class-D among the compared variants while retaining semantic similarity.

Table 4: Ablation of the STELLAR inference operators.

Setting	Sym.	Rel.	HGC	CLIP Sim ↑	CLIP Dir ↓	SDI-CD ↓	LoD-CD ↓	GD ↓	1-GD ↓	FPD ↓	Class-D ↓
Original	×	×	×	.2600	.990	.6742	.2688	**.340**	**.070**	33.64	.0500
STELLAR w/o Rel.	✓	×	✓	.2602	.989	.6548	.2637	.343	.071	33.42	.0499
STELLAR w/o HGC	✓	✓	×	.2604	.988	.6519	.2631	.345	.071	33.35	.0498
STELLAR w/o Sym.	×	n/a	✓	.2603	.989	.6635	.2648	.343	.071	33.39	.0499
STELLAR	✓	✓	✓	**.2607**	**.987**	**.6422**	**.2591**	.348	**.070**	**33.17**	**.0497**

A.2 CATEGORY-LEVEL RESULTS

The category-level comparison uses the same paired protocol as the main evaluation. STELLAR improves the structural metrics on tables and lamps while maintaining or slightly improving semantic similarity.

Table 5: Category-level paired results. Lower is better except CLIP Sim.

Category	Method	CLIP Sim ↑	CLIP Dir ↓	SDI-CD ↓	LoD-CD ↓	GD ↓	1-GD ↓	FPD ↓	Class-D ↓
Chair	BlendedPC	.2600	.990	.6742	.2688	**.340**	**.070**	33.64	.0500
Chair	STELLAR	**.2607**	**.987**	**.6422**	**.2591**	.348	**.070**	**33.17**	**.0497**
Table	BlendedPC	.2363	1.005	.0724	.0308	.365	.082	42.10	.0730
Table	STELLAR	**.2371**	**.997**	**.0707**	**.0302**	**.360**	**.080**	**41.40**	**.0710**
Lamp	BlendedPC	.2290	1.015	.0630	.0277	.355	.078	38.70	.0670
Lamp	STELLAR	**.2297**	**1.006**	**.0618**	**.0272**	**.350**	**.076**	**37.90**	**.0650**

A.3 Part-Level Results

Table 6 breaks down the paired results by edited part. The structural guidance is applied only to editable coordinates; the protected region continues to use the reconstruction branch.

Table 6: Part-level paired results. Lower is better except CLIP Sim.

Category	Edited part	Method	CLIP Sim ↑	SDI-CD ↓	LoD-CD ↓	Full-CD ↓
Chair	Leg	BlendedPC	.2295	.0685	.0281	.0178
Chair	Leg	STELLAR	**.2302**	**.0667**	**.0275**	**.0176**
Chair	Seat	BlendedPC	.2470	.0612	.0258	.0165
Chair	Seat	STELLAR	**.2474**	**.0608**	**.0254**	**.0164**
Chair	Back	BlendedPC	.2520	.0588	.0246	.0161
Chair	Back	STELLAR	**.2522**	**.0581**	**.0241**	**.0160**
Chair	Armrest	BlendedPC	.2415	.0647	.0274	.0172
Chair	Armrest	STELLAR	**.2421**	**.0630**	**.0265**	**.0170**
Table	Leg	BlendedPC	.2340	.0745	.0315	.0188
Table	Leg	STELLAR	**.2348**	**.0728**	**.0307**	**.0185**
Lamp	Stem/base	BlendedPC	.2280	.0650	.0280	.0175
Lamp	Stem/base	STELLAR	**.2286**	**.0638**	**.0274**	**.0173**

A.4 Inference Efficiency

Efficiency is measured after one warm-up run using the same fixed sampling protocol. STELLAR adds a small runtime cost relative to BlendedPC while keeping nearly identical peak memory use.

Table 7: Inference efficiency and robustness. Time and memory are averaged over valid generated samples.

Method	Time (s) ↓	Memory (GB) ↓	Valid (%) ↑
ChangeIt3D	19.362	4.659	99
SPICE-E	**15.421**	**2.376**	92
BlendedPC	17.782	3.429	**100**
STELLAR	18.255	3.431	**100**

B Evaluation Protocol and Experimental Settings

These settings specify the paired comparison and the reported parameter configuration; they are not additional outcome claims.

B.1 Experimental Protocol

We evaluate the same self-trained localized diffusion carrier under two inference settings. In the Original setting, all structural guidance is disabled. In the STELLAR setting, part-conditioned symmetry, position-aware reliability, and hierarchical geometric consistency are enabled. Both settings use the same checkpoint, input cloud, instruction, mask, random seed, and sampling steps.

Table 8: Inference configurations used in the paired comparison.

Setting	Symmetry	Reliability / HGC
Original	Disabled	Disabled
STELLAR	Enabled	Position-aware / Enabled

B.2 Category–Part Prior Specification

The category–part prior is selected from a fixed rule table before inference. It is not selected by inspecting a generated output or a target shape. For a category–part pair without a reliable repeated relation, STELLAR keeps the identity transformation and disables the symmetry correction. The source reliability field and HGC operator can still be used in that case.

Table 9: Category–part structural prior specification.

Category	Edited part	Structural relation	Prior type
Chair	Leg	Bilateral repetition	Reflection
Chair	Armrest	Bilateral repetition	Reflection
Chair	Back	Approximate bilateral relation	Reflection / rotation
Chair	Seat	Local continuity	No forced symmetry
Table	Leg / support	Repeated supports	Reflection
Table	Top	Local continuity	No forced symmetry
Lamp	Stem / base / shade	Axial or bilateral relation	Reflection / rotation
Airplane	Wing / tail	Bilateral relation	Reflection / rotation

B.3 Evaluation Metrics

For two point clouds $\mathcal{X}$ and $\mathcal{Y}$, the symmetric Chamfer distance is

$$\mathrm{CD}(\mathcal{X},\mathcal{Y}) = \frac{1}{|\mathcal{X}|}\sum_{\mathbf{x}\in\mathcal{X}}\min_{\mathbf{y}\in\mathcal{Y}}\|\mathbf{x}-\mathbf{y}\|_2^2 + \frac{1}{|\mathcal{Y}|}\sum_{\mathbf{y}\in\mathcal{Y}}\min_{\mathbf{x}\in\mathcal{X}}\|\mathbf{y}-\mathbf{x}\|_2^2. \tag{13}$$

CLIP Sim measures semantic similarity and CLIP Dir measures semantic directional error. SDI-CD measures structural deviation, while LoD-CD measures geometric discrepancy across the evaluation resolutions $\{256, 512, 1024\}$. GD, 1-GD, and Full-CD measure geometric distortion and source-shape preservation. FPD and Class-D measure distributional and category-level fidelity.

Table 10: Summary of evaluation metrics and preferred directions.

Metric	Evaluation aspect	Direction
CLIP Sim	Semantic similarity	Higher
CLIP Dir	Semantic directional error	Lower
SDI-CD	Structural deviation	Lower
LoD-CD	Multi-scale geometric discrepancy	Lower
GD	Geometric distortion	Lower
1-GD	Identity-preservation discrepancy	Lower
Full-CD	Full-shape source deviation	Lower
FPD	Distributional discrepancy	Lower
Class-D	Category-level distortion	Lower

B.4 Detailed Metric Protocol

All compared outputs are represented in the same normalized coordinate system and evaluated with the same point-count convention. For two point clouds $\mathcal{X}$ and $\mathcal{Y}$, the symmetric Chamfer distance in Section B.3 is used as the common geometric primitive. SDI-CD evaluates the departure of the editable subset from its selected category–part relation. LoD-CD averages geometric discrepancies at the evaluation resolutions $\{256, 512, 1024\}$; these evaluation resolutions are separate from the HGC implementation scales $\{16, 64, 128\}$.

Let $\mathcal{P}_s(\cdot)$ denote the point-cloud representation at evaluation resolution s. The scale-specific LoD discrepancy is

$$D_s = \mathrm{CD}\big(\mathcal{P}_s(\widehat{\mathcal{X}}), \mathcal{P}_s(\mathcal{X}^{\mathrm{ref}})\big), \tag{14}$$

and the reported multi-scale value is

$$\text{LoD-CD} = \frac{1}{3} \sum_{s \in \{256,512,1024\}} D_s. \tag{15}$$

Here $\mathcal{X}^{\text{ref}}$ follows the reference construction of the corresponding evaluation protocol. Full-CD compares the final output to the source shape, whereas GD and 1-GD follow the ShapeTalk metric implementation. FPD compares generated and reference distributions in the evaluation feature space, and Class-D measures category-level distortion.

The Original and STELLAR settings are evaluated as paired outputs. For any metric m, the per-instance difference is

$$\delta_i^{(m)} = m_i^{\text{STELLAR}} - m_i^{\text{Original}}. \tag{16}$$

Thus, a negative difference indicates an improvement for lower-is-better metrics, while a positive difference indicates an improvement for CLIP Sim. Only valid outputs are included in an aggregate metric; invalid outputs remain recorded in the run-status file and contribute to the reported valid-output rate.

B.5 Training and Inference Configuration

Table 11 lists the training configuration of the localized diffusion carrier. Table 12 lists the frozen STELLAR inference configuration used for the reported results. The three HGC scales are implementation levels and are distinct from the LoD-CD evaluation resolutions $\{256, 512, 1024\}$.

Table 11: Training configuration of the localized diffusion carrier.

Parameter	Value
Physical batch size	32
Gradient accumulation	2
Effective batch size	64
Learning rate	1×10^{-4}
Copy probability	0.1
Conditional-drop probability	0.5
Symmetry-loss weight	0
Random seed	2026
Checkpoint selection	Validation-selected checkpoint

Table 12: STELLAR inference configuration.

Parameter	Value
Base / upsampling steps	64 / 64
Symmetry guidance weight	.06
Final correction weight	.015
Reliability bandwidth σ	.15
Minimum reliability weight	.25
Reliability exponent γ	1
HGC weight	.005
HGC scales	$\{16, 64, 128\}$
Maximum displacement	.04
Protected bypass / coordinate blending	Enabled

B.6 Asymmetric Instructions and Prior Compatibility

Symmetry guidance is not appropriate for an instruction that explicitly requests a one-sided deformation, removal of a single component, or another inherently asymmetric change. The prior table therefore acts as a configuration rule rather than a universal constraint. For an incompatible category–part pair, the symmetry operator is disabled while the protected reconstruction branch, the position-aware reliability field, and the HGC operator retain their documented roles.

We assess prior compatibility by keeping the source point cloud, instruction, mask, checkpoint, seed, and sampling budget fixed across the following settings: the selected category–part prior, no symmetry prior, an incorrect prior, and a perturbed symmetry axis. This protocol separates an error caused by a mismatched structural assumption from an error caused by the diffusion carrier or the input mask. The comparison is qualitative unless paired per-sample metrics for each setting are available.

Table 13: Prior-compatibility settings used for supplementary qualitative analysis.

Setting	Purpose
Selected prior	Apply the registered rule for the category–part pair.
No symmetry prior	Isolate the contribution of symmetry correction.
Incorrect prior	Test sensitivity to an incompatible structural relation.
Perturbed axis	Test sensitivity to a shifted or misaligned symmetry axis.
Asymmetric instruction	Disable symmetry when the instruction conflicts with it.

B.7 Parameter Selection Protocol

All guidance parameters are selected on a validation protocol and then fixed for the reported test evaluation. The selection procedure prioritizes structural improvement, then verifies that semantic fidelity and source-shape preservation do not show a material regression. No test example is used to choose a parameter value.

Table 14: Candidate validation ranges and the frozen reported configuration.

Parameter	Candidate values	Reported value
Symmetry weight	.04, .06, .08, .12	.06
Final symmetry weight	.01, .015, .02	.015
Reliability bandwidth σ	.10, .15, .20	.15
Minimum reliability weight	.10, .25, .40	.25
Reliability exponent γ	.5, 1, 2	1
HGC weight	0, .003, .005, .01	.005
Maximum displacement	.02, .04, .06	.04

C Additional Method and Implementation Details

This section gives the full formulation of the localized diffusion carrier, the protected reconstruction branch, and the three structural operators.

C.1 Notation and Problem Formulation

Let the source point cloud be

$$\mathcal{X}^{\mathrm{src}} = \{\mathbf{x}_i^{\mathrm{src}}\}_{i=1}^{N} \in \mathbb{R}^{N\times 3}, \tag{17}$$

where $\mathbf{x}_i^{\mathrm{src}} \in \mathbb{R}^3$ is the coordinate of the i-th point. Given an instruction c, an object category y, a target part p, and a binary editing mask $\mathbf{M} \in \{0,1\}^N$, we define

$$\mathcal{I}_e = \{i \mid M_i = 1\}, \qquad \mathcal{I}_k = \{i \mid M_i = 0\}, \tag{18}$$

where $\mathcal{I}_e$ and $\mathcal{I}_k$ denote the editable and protected index sets, respectively. The associated diagonal mask operators are

$$\mathbf{D}_M = \mathrm{diag}(M_1, \ldots, M_N), \qquad \mathbf{D}_{\bar{M}} = \mathbf{I}_N - \mathbf{D}_M. \tag{19}$$

For an output point cloud $\widehat{\mathcal{X}}$, the two regions are

$$\widehat{\mathcal{X}}_e = \mathbf{D}_M \widehat{\mathcal{X}}, \qquad \widehat{\mathcal{X}}_k = \mathbf{D}_{\bar{M}} \widehat{\mathcal{X}}. \tag{20}$$

The localized editing objective is to generate

$$\widehat{\mathcal{X}} = \mathcal{F}_\theta(\mathcal{X}^{\mathrm{src}}, c, y, p, \mathbf{M}) \tag{21}$$

such that the editable region follows c, the protected region remains close to the source, and the edited part preserves the structural relation specified by the category–part prior $\mathcal{G}_{y,p}$.

For analysis, these requirements can be summarized as

$$\widehat{\mathcal{X}}^* = \arg\min_{\widetilde{\mathcal{X}}} \Big[\lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}}(\widetilde{\mathcal{X}}_e, c) + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}(\widetilde{\mathcal{X}}_k, \mathcal{X}^{\mathrm{src}}_k) + \lambda_{\mathrm{str}}\mathcal{L}_{\mathrm{str}}(\widetilde{\mathcal{X}}_e, \mathcal{G}_{y,p})\Big]. \tag{22}$$

This expression describes the desired properties and is not used as an additional training loss in our implementation.

C.2 Localized Diffusion Formulation

Given a clean training point cloud $\mathcal{X}_0$, the forward process is

$$\mathbf{X}_t = \sqrt{\bar{\alpha}_t}\mathbf{X}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{23}$$

where $\alpha_t = 1 - \beta_t$ and

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s. \tag{24}$$

The localized denoiser receives the noisy point cloud, the text condition, the editing mask, and the timestep:

$$\widehat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_\theta(\mathbf{X}_t, c, \mathbf{M}, t). \tag{25}$$

With conditional dropout probability p_{drop}, the condition is

$$\widetilde{c} = \begin{cases} \varnothing, & \text{with probability } p_{\mathrm{drop}}, \\ c, & \text{otherwise.} \end{cases} \tag{26}$$

The backbone is optimized with the standard noise-prediction objective

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{\mathbf{X}_0, c, \mathbf{M}, t, \boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{X}_t, \widetilde{c}, \mathbf{M}, t)\|_2^2\right]. \tag{27}$$

The clean estimate recovered at reverse step t is

$$\widehat{\mathbf{X}}_0^{(t)} = \frac{\mathbf{X}_t - \sqrt{1-\bar{\alpha}_t}\,\widehat{\boldsymbol{\epsilon}}_t}{\sqrt{\bar{\alpha}_t}}. \tag{28}$$

Its editable proposal is

$$\widehat{\mathbf{X}}_{0,e}^{(t)} = \mathbf{D}_M \widehat{\mathbf{X}}_0^{(t)}. \tag{29}$$

The total training objective used in all reported experiments is therefore

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{diff}}. \tag{30}$$

The STELLAR operators are applied to the clean estimate during reverse diffusion and do not introduce trainable parameters into the backbone.

C.3 Protected Reconstruction and Coordinate Blending

To preserve the non-edited geometry, reverse diffusion maintains an editing branch and a reconstruction branch initialized with shared noise:

$$\mathbf{X}_T^{\mathrm{edit}} = \mathbf{X}_T^{\mathrm{rec}} = \sqrt{\bar{\alpha}_T}\mathcal{X}^{\mathrm{src}} + \sqrt{1-\bar{\alpha}_T}\,\boldsymbol{\epsilon}. \tag{31}$$

At step t, the editing branch produces a text-conditioned proposal

$$\widehat{\mathbf{X}}_{0,\mathrm{edit}}^{(t)} = \mathcal{D}_\theta(\mathbf{X}_t^{\mathrm{edit}}, c, \mathbf{M}, t), \tag{32}$$

while the reconstruction branch produces a source-preserving estimate

$$\widehat{\mathbf{X}}_{0,\mathrm{rec}}^{(t)} = \mathcal{R}_\theta(\mathbf{X}_t^{\mathrm{rec}}, \mathbf{M}, t). \tag{33}$$

Only the editable part of the editing proposal is passed to STELLAR:

$$\widehat{\mathbf{X}}_{0,e}^{(t)} = \mathbf{D}_M \widehat{\mathbf{X}}_{0,\text{edit}}^{(t)}. \tag{34}$$

After structural refinement, the two branches are combined by

$$\mathbf{X}_{t-1} = \mathbf{D}_M \widetilde{\mathbf{X}}_{t-1}^{\text{edit}} + \mathbf{D}_{\bar{M}} \mathbf{X}_{t-1}^{\text{rec}}. \tag{35}$$

Equivalently,

$$\mathbf{x}_{i,t-1} = \begin{cases} \widetilde{\mathbf{x}}_{i,t-1}^{\text{edit}}, & i \in \mathcal{I}_e, \\ \mathbf{x}_{i,t-1}^{\text{rec}}, & i \in \mathcal{I}_k. \end{cases} \tag{36}$$

This protected-coordinate bypass ensures that STELLAR directly corrects only the editable region.

C.4 Part-Conditioned Symmetry Projection

For each category–part pair (y, p), we define a finite set of candidate reflections or rotations

$$\mathcal{G}_{y,p} = \{g_1, \ldots, g_{K_g}\}. \tag{37}$$

For an editable point $\widehat{\mathbf{x}}_i^{(t)}$, each candidate gives

$$\mathbf{q}_{i,k}^{(t)} = g_k(\widehat{\mathbf{x}}_i^{(t)}). \tag{38}$$

Because point clouds are unordered, correspondence is obtained by nearest neighbor matching:

$$\pi_k(i) = \arg\min_{j \in \mathcal{I}_e} \left\| \mathbf{q}_{i,k}^{(t)} - \widehat{\mathbf{x}}_j^{(t)} \right\|_2. \tag{39}$$

The candidate with the smallest residual is selected:

$$k^*(i) = \arg\min_k \left\| g_k(\widehat{\mathbf{x}}_i^{(t)}) - \widehat{\mathbf{x}}_{\pi_k(i)}^{(t)} \right\|_2. \tag{40}$$

The symmetry displacement is

$$\Delta_{\text{sym}}^{(t)}(\widehat{\mathbf{x}}_i^{(t)}) = \widehat{\mathbf{x}}_{\pi_{k^*(i)}(i)}^{(t)} - g_{k^*(i)}(\widehat{\mathbf{x}}_i^{(t)}). \tag{41}$$

To prevent a single unreliable correspondence from producing a large jump, we use norm clipping:

$$\widetilde{\Delta}_{\text{sym}}^{(t)} = \Delta_{\text{sym}}^{(t)} \min\left(1, \frac{\delta_{\text{sym}}}{\|\Delta_{\text{sym}}^{(t)}\|_2 + \varepsilon}\right). \tag{42}$$

The corresponding structural correction is

$$\mathbf{G}_{\text{sym}}^{(t)}(\mathbf{x}_i) = \lambda_{\text{sym}}(t)\, m_i^{(t)}\, \widetilde{\Delta}_{\text{sym}}^{(t)}(\mathbf{x}_i), qquadi \in \mathcal{I}_e. \tag{43}$$

For protected points, $\mathbf{G}_{\text{sym}}^{(t)}(\mathbf{x}_i) = \mathbf{0}$.

C.5 Dynamic Spatial Reliability Calibration

The reliability field reduces symmetry corrections near fragile edit–keep interfaces. For each editable point, we compute its distance to the protected region:

$$d_i = \min_{j \in \mathcal{I}_k} \|\mathbf{x}_i - \mathbf{x}_j\|_2. \tag{44}$$

Let d_{box} be the diagonal length of the source bounding box. The normalized distance is

$$\bar{d}_i = \frac{d_i}{d_{\text{box}} + \varepsilon}. \tag{45}$$

The position-aware reliability weight is

$$m_i = m_{\min} + (1 - m_{\min}) \left[1 - \exp\left(-\frac{\bar{d}_i^2}{2\sigma^2}\right)\right]^{\gamma}. \tag{46}$$

Here $m_{\min}$ is the minimum weight near the protected boundary, σ controls the spatial transition bandwidth, and γ controls the shape of the transition. The gated symmetry displacement becomes

$$\Delta_{\text{sym}}^{\text{rel},(t)}(\mathbf{x}_i) = m_i \widetilde{\Delta}_{\text{sym}}^{(t)}(\mathbf{x}_i). \tag{47}$$

The field is computed from the source geometry and reused during sampling; it is not a learned auxiliary network.

C.6 Hierarchical Geometric Consistency

The hierarchical geometric consistency operator suppresses common structural drift at multiple geometric scales. Let

$$\mathcal{S}_{\mathrm{hgc}} = \{16, 64, 128\} \tag{48}$$

denote the implementation hierarchy. For each level s, farthest-point sampling or the corresponding fixed cluster assignment partitions the editable points into clusters $\{\mathcal{C}_\ell^{(s)}\}_{\ell=1}^{K_s}$. The proposal and reference cluster means are

$$\boldsymbol{\mu}_\ell^{(s)} = \frac{1}{|\mathcal{C}_\ell^{(s)}|} \sum_{i \in \mathcal{C}_\ell^{(s)}} \mathbf{x}_i, qquad \widehat{\boldsymbol{\mu}}_\ell^{(s)} = \frac{1}{|\mathcal{C}_\ell^{(s)}|} \sum_{i \in \mathcal{C}_\ell^{(s)}} \widehat{\mathbf{x}}_i. \tag{49}$$

The level-specific residual is

$$\Delta_s(\mathbf{x}_i) = \boldsymbol{\mu}_{\ell(i)}^{(s)} - \widehat{\boldsymbol{\mu}}_{\ell(i)}^{(s)}, qquad i \in \mathcal{C}_{\ell(i)}^{(s)}. \tag{50}$$

To avoid excessive displacement, the residual is clipped by

$$\mathrm{clip}_{\delta_s}(\Delta) = \Delta \min\left(1, \frac{\delta_s}{\|\Delta\|_2 + \varepsilon}\right). \tag{51}$$

The aggregated hierarchical correction is

$$\Delta_{\mathrm{hgc}}(\mathbf{x}_i) = \sum_{s \in \mathcal{S}_{\mathrm{hgc}}} \alpha_s \,\mathrm{clip}_{\delta_s}\left(\Delta_s(\mathbf{x}_i)\right), \qquad \sum_s \alpha_s = 1. \tag{52}$$

The implementation hierarchy $\{16, 64, 128\}$ is distinct from the evaluation resolutions used by the LoD-CD metric, which are $\{256, 512, 1024\}$.

C.7 Static and Dynamic Quantities

The source point cloud, the edit mask, the category–part rule, the reliability field inputs, and the HGC FPS labels are fixed before reverse diffusion starts. In particular, the reliability field is derived from the source geometry and the edit–keep boundary; it is not a learned auxiliary network and is not re-estimated from a generated output.

By contrast, the editable clean proposal, nearest-neighbor correspondences, symmetry displacement, cluster means, and HGC residuals change at each reverse step. This distinction ensures that the structural prior stays stable while the corrective displacement responds to the current diffusion proposal. The protected coordinates do not receive symmetry or HGC correction and reach the final output through the reconstruction trajectory and coordinate blending.

C.8 Guidance Fusion and Reverse-Diffusion Update

At reverse step t, the three corrections are combined as

$$\mathbf{G}^{(t)}(\mathbf{x}_i) = \lambda_{\mathrm{sym}}(t)\Delta_{\mathrm{sym}}^{\mathrm{rel},(t)}(\mathbf{x}_i) + \lambda_{\mathrm{hgc}}(t)\Delta_{\mathrm{hgc}}^{(t)}(\mathbf{x}_i). \tag{53}$$

The corrected clean estimate is

$$\widetilde{\mathbf{x}}_{i,0}^{(t)} = \widehat{\mathbf{x}}_{i,0}^{(t)} + \mathbf{G}^{(t)}(\mathbf{x}_i). \tag{54}$$

The correction is bounded by the maximum displacement $\delta_{\max}$:

$$\widetilde{\mathbf{x}}_{i,0}^{(t)} = \widehat{\mathbf{x}}_{i,0}^{(t)} + \mathbf{G}^{(t)}(\mathbf{x}_i) \min\left(1, \frac{\delta_{\max}}{\|\mathbf{G}^{(t)}(\mathbf{x}_i)\|_2 + \varepsilon}\right). \tag{55}$$

The numerical sampler then performs the transition

$$\widetilde{\mathbf{X}}_{t-1}^{\mathrm{edit}} = \Phi_t\left(\mathbf{X}_t^{\mathrm{edit}}, \widetilde{\mathbf{X}}_{0,e}^{(t)}, t\right). \tag{56}$$

Finally, mask-based coordinate blending combines the edited and protected branches as described in Section C.3.

C.9 Inference Procedure

The complete inference procedure is summarized below.

Input: source cloud $\mathcal{X}^{\mathrm{src}}$, instruction c, mask $\mathbf{M}$, category–part prior $\mathcal{G}_{y,p}$, denoiser ϵ_θ, sampler Φ, and reverse steps T.
Initialize: sample shared noise and initialize editing and reconstruction trajectories.
For $t = T, \ldots, 1$:
Predict the text-conditioned clean estimate.
Extract the editable subset using $\mathbf{M}$.
Apply the category–part symmetry projection.
Compute the position-aware reliability weights m_i.
Apply the reliability-weighted symmetry correction.
Estimate and aggregate hierarchical geometric residuals.
Fuse the bounded corrections into the clean editable estimate.
Apply the numerical sampler update to the edited branch.
Update the reconstruction branch.
Blend edited coordinates with protected reconstruction coordinates.
Return: the final blended point cloud $\widehat{\mathcal{X}}$.

C.10 Training and Inference-Time Separation

The training and inference roles are summarized in Table 15. The diffusion carrier learns text-guided localized editing from the diffusion objective, while STELLAR controls the reverse trajectory through deterministic structural operators.

Table 15: Separation between diffusion-carrier training and STELLAR inference.

Stage	Main operation	Role
Training	Noise prediction	Learn localized text-guided editing
Inference	Symmetry projection	Establish part-level correspondence
Inference	Reliability calibration	Reduce boundary-sensitive corrections
Inference	Hierarchical consistency	Suppress multi-scale structural drift
Inference	Coordinate blending	Preserve protected coordinates

C.11 Implementation and Reproducibility

All experiments use the same ShapeTalk preprocessing, normalized point-cloud coordinate convention, and fixed sample manifests for paired comparisons. For the reported baseline training, the diffusion carrier uses

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{diff}}, \tag{57}$$

with batch size, gradient accumulation, learning rate, seed, and checkpoint selection recorded in the corresponding experiment log. The inference comparison uses the validation-selected checkpoint rather than automatically using the final epoch.

The Original setting uses

$$\texttt{symmetry_type=none}, \quad \texttt{lod_weight=0}, \tag{58}$$

whereas STELLAR uses the registered symmetry, position-aware reliability, and hierarchical consistency arguments. Both settings use identical input rows, prompts, masks, seeds, and diffusion steps.

For reproducibility, each run records the checkpoint path, dataset split, sample manifest, complete command, GPU index, seed, output directory, per-case return code, elapsed time, and output-file validity. Invalid outputs are retained in the status file and are never silently counted as successful samples.

D Qualitative Visualizations and Limitations

D.1 Failure Cases

The following cases should be reported as limitations rather than removed from the evaluation: an inherently asymmetric instruction, an incorrect category–part prior, an edit touching a connected boundary, sparse input points, a large deformation, or an invalid part mask.

Table 16: Failure-case taxonomy for qualitative analysis.

Case	Interpretation
Asymmetric instruction	Symmetry may conflict with the requested one-sided edit.
Incorrect prior	The selected transformation does not match the target part.
Connected boundary	Strong correction may affect a fragile attachment region.
Sparse point density	Correspondence and cluster residuals become unreliable.
Large deformation	Bounded refinement may not fully recover the target geometry.
Invalid mask	The editable region does not accurately identify the target part.

D.2 Additional Qualitative Results

We provide additional category-level qualitative results for chairs, tables, and lamps. The figures use the same color convention and compare the input, the edited result, and the corresponding baseline outputs where available.

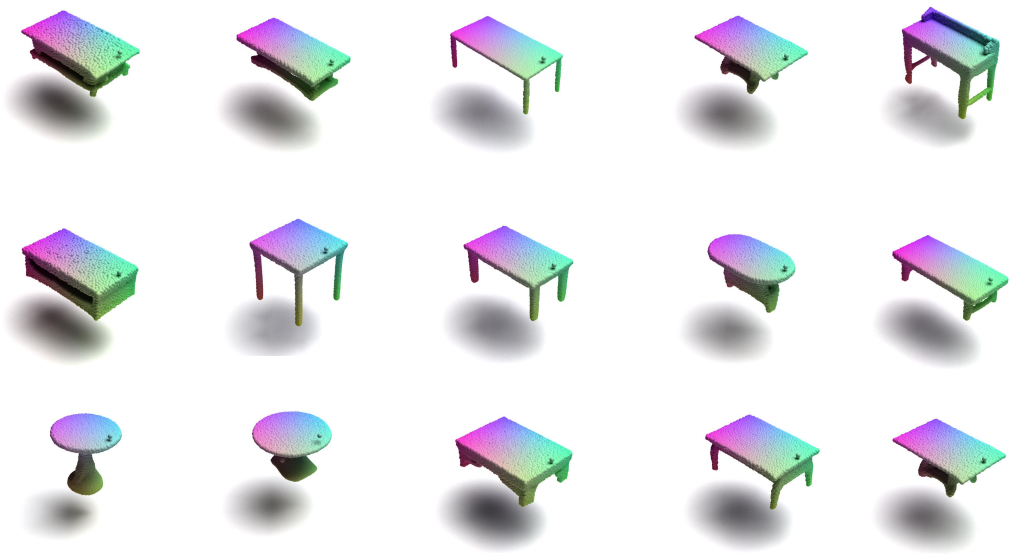

Figure 4: Supplementary qualitative visualization 1.

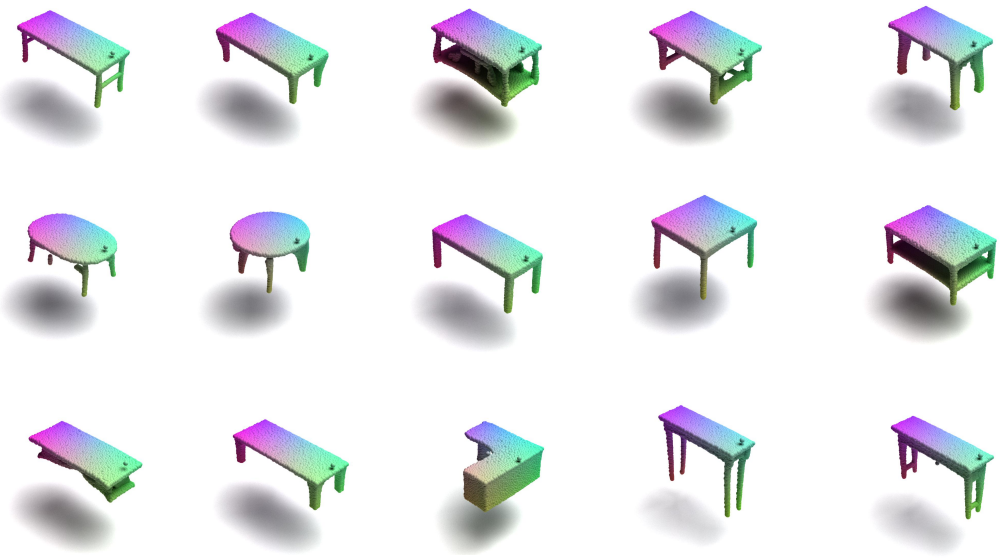

Figure 5: Supplementary qualitative visualization 2.



Figure 6: Supplementary qualitative visualization 3.

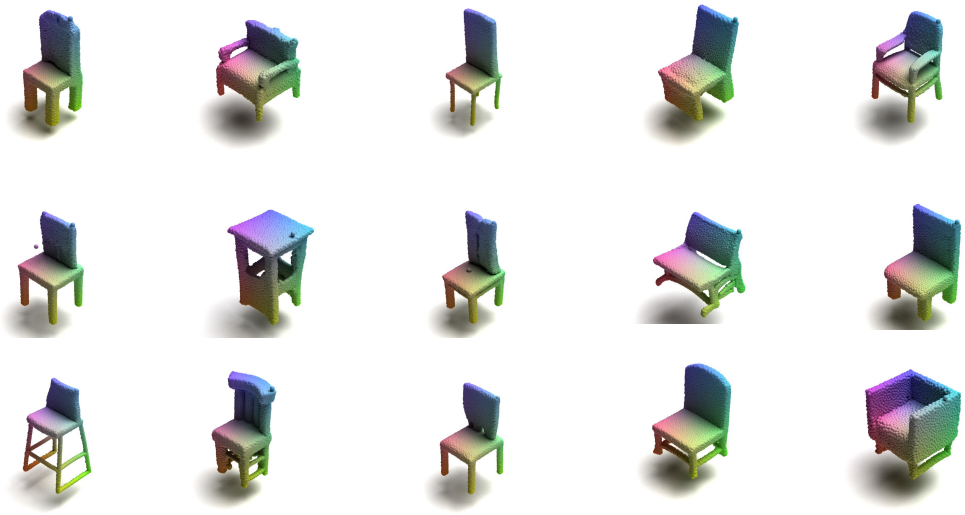

Figure 7: Supplementary qualitative visualization 4.

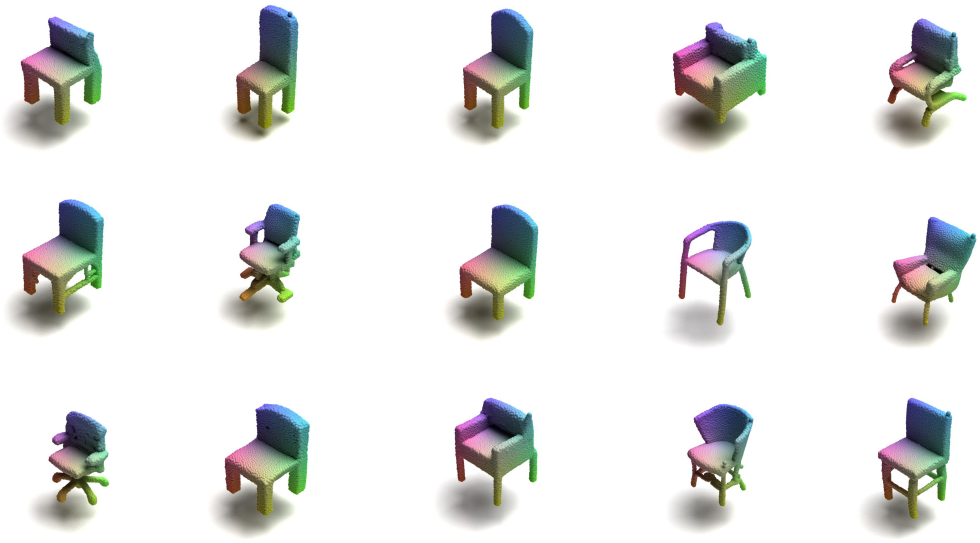

Figure 8: Supplementary qualitative visualization 5.

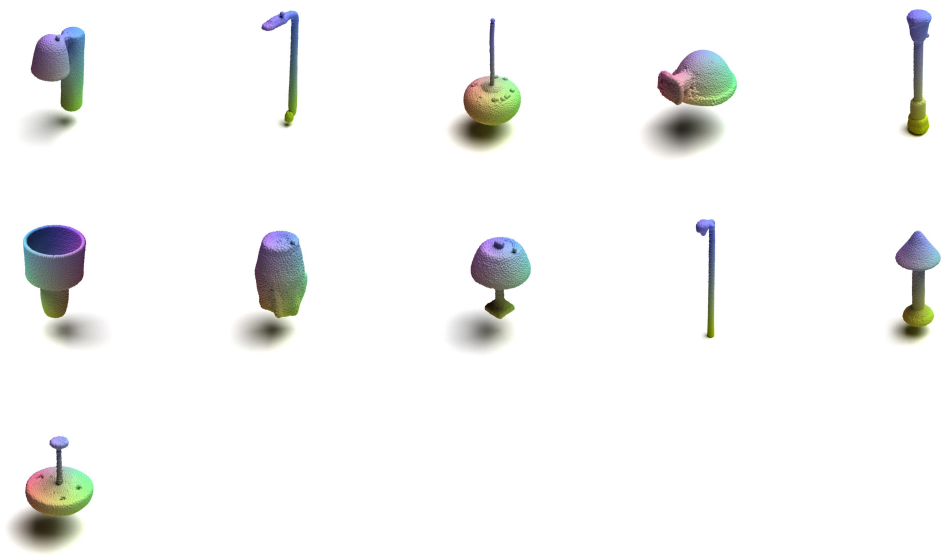

Figure 9: Supplementary qualitative visualization 6.