Intrinsic-Preserving Schrödinger Bridge for Direct Part-Aware 3D Generation

Qitong Yang
Xidian University

Mingtao Feng*
Xidian University

Zijie Wu*
The University of Western Australia

Jie Feng
Xidian University

Weisheng Dong
Xidian University

Ajmal Mian
The University of Western Australia

Abstract

Controllable 3D generation is a fundamental challenge in computer vision. Existing classifier-free guidance methods often yield suboptimal alignment with the given conditions, while Schr"odinger Bridge approaches overlook intrinsic geometric discrepancies across multimodal manifolds, leading off-manifold trajectories and semantic misalignment. To address these issues, we propose Intrinsic-Preserving Schrödinger Bridge (IPSB), a novel framework for direct part-aware 3D generation that leverages cross-modal intrinsic consistency. Specifically, we introduce Hierarchical Part-level Cross-modal Encoder that constructs hierarchical part-level representations through probability-based soft coverage, explicitly constraining intrinsic to establish semantically consistent part graphs across modalities. Additionally, our IPSB incorporates asymmetric GW-inspired relational regularization, which strictly preserves the intrinsic semantic structure in the conditional manifold while allowing sufficient geometric flexibility in the 3D target manifold. Based on our IPSB framework, we devise Compositionally Controllable 3D Generation based on part-wise intervention inference, enabling independent replacement or resampling of target part nodes while fixing the trajectories of remaining components, thus achieving part-level controllable editing without full regeneration. Extensive experiments demonstrate that our superiority, efficiency and generality.

1 Introduction

Conditional 3D shape generation synthesizes 3D geometry from multimodal inputs such as text or images, and is fundamental to vision and graphics, with applications in digital asset creation [59, 61, 39, 62], virtual reality [41, 55], autonomous driving simulation [38, 63], and robotics [60, 2, 58].

Current methods mainly uses diffusion [48, 65, 56] or flow-based models [36, 1, 64, 32] with Classifier-Free Guidance (CFG) [20] to steer sampling toward condition-aligned outputs via global trajectory perturbation. **1)** However, CFG operates at the distribution level rather than part-level alignment, lacking explicit part-level alignment: it enforces global consistency without part-to-part correspondence (see Fig. 1(a)). This lack of part-level transport cost prevents fine-grained semantic alignment and compositional control, regardless of guidance scale. Although part-aware methods [67, 7, 8, 61, 62] improve via attention-based assignment but remain constrained by the CFG paradigm. Recent advances in Schrödinger Bridge (SB) theory [11, 49] provide principled optimal

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

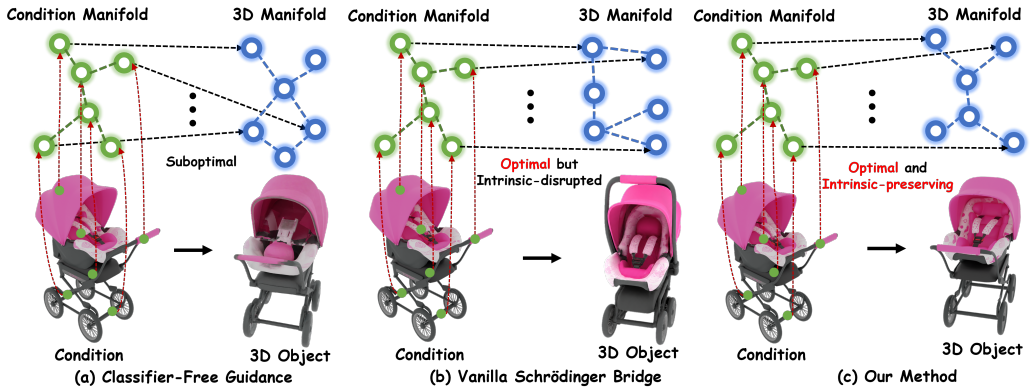

Figure 1: Comparison of the mechanisms of CFG, vanilla SB, and our method.

transport formulations for generative modeling beyond CFG, with applications in 3D generation such as TraCe [33], P2P-Bridge [54], and BridgeShape [26]. However, these SB-based methods define transport cost extrinsically in a shared embedding space, relying on pointwise comparisons. **2)** When source and target modalities lie on intrinsically different manifolds, such an extrinsic cost fails to capture meaningful semantic correspondence, leading transport paths to drift off the target manifold and disrupting part-level relational structure (see Fig. 1(b)). **3)** Neither CFG nor existing SB methods explicitly model the intrinsic of cross-modal feature spaces, preventing intrinsically structure-preserving inter-modal mappings. The identified limitations share a common root: the lack of modeling intrinsic topology in cross-modal transport. Recent Platonic [22] and Aristotelian [16] hypotheses provide evidence for cross-modal optimal transport: deep networks learn modality-specific metric structures, they preserve consistent intrinsic topology. If samples A and B are mutual nearest neighbors in one modality, they tend to remain so in another. This suggests transport should be defined by intrinsic relational structure rather than extrinsic pointwise distances. Motivated by this insight, we pose a *question*: *can we exploit cross-modal intrinsic consistency to establish direct optimal transport mappings between modalities, thereby enabling semantically-aware part-level 3D generation?*

To address limitations of CFG-based and SB approaches while exploiting cross-modal intrinsic, we propose Intrinsic-Preserving Schrödinger Bridge (IPSB), a framework for intrinsic-preserving cross-modal optimal transport via part-level cross-modal encoding and asymmetric Gromov-Wasserstein [42](GW)-inspired relational regularization (see Fig. 1(c)). **1)** To build semantically consistent representations for transport, we introduce a Hierarchical Part-level Cross-modal Encoder that uses a Riemannian Gaussian mixture model for adaptive semantic coverage while explicitly constraining graph topology to follow the intrinsic structure of conditioning semantics. This design aligns part-level graphs with semantic hierarchies and preserves consistency across different granularities, forming structured foundations for subsequent optimal transport. **2)** To constrain transport while respecting cross-modal directional asymmetry, we propose Intrinsic-Preserving Schrödinger Bridge (IPSB), which optimal transport from pointwise to intrinsic relational form. Here, intrinsic refers to the pairwise relational structure among part-level embeddings, represented by graph-geodesic distances on the part graph rather than their absolute coordinates in the shared embedding space. By using asymmetric GW-inspired regularization on modality-specific part graphs, IPSB replaces ambient distances and jointly resolves part-level misalignment in CFG and manifold-level inconsistency in standard SB within a unified framework. Specifically, the regularization reflects cross-modal directionality by assigning asymmetric weights: intrinsic violations in the conditional space are penalized as unacceptable inconsistencies, while those in the 3D space are allowed as limited geometric deformation. **3)** Building on the Hierarchical Part-level Cross-modal Encoder and IPSB, we introduce Compositionally Controllable 3D Generation via part-wise intervention inference mechanism that enables independent substitution or resampling of target parts while keeping other transport trajectories fixed, supporting localized editing without full regeneration, which is not achievable in CFG-based or existing SB methods. Our main contributions are:

- We propose a Hierarchical Part-level Cross-modal Encoder to achieve adaptive semantic soft coverage with explicit topological constraints, constructing part graphs that provide semantically consistent feature foundations for cross-modal optimal transport.

- We introduce Intrinsic-Preserving Schrödinger Bridge with asymmetric GW-inspired relational regularization, aligning transport regularization with the intrinsic directionality of cross-modal generation while preserving both data alignment and intrinsic.
- We develop Compositionally Controllable 3D Generation based on our IPSB, supporting independent intervention and resampling of individual parts to achieve compositional control without holistic regeneration.

Extensive experiments on image-to-3D generation, text-to-3D generation, 3D-to-3D generation, part-aware editing, and scene generation demonstrate the effectiveness, superiority, and generality of our method.

2 Related Work

3D Generation. Early text-to-3D methods use pretrained 2D diffusion with Score Distillation Sampling (SDS), including DreamFusion [43], DreamGaussian [52], and MVDream [48], but suffer from slow optimization and limited fidelity, motivating direct conditional generation. Feed-forward models such as LGM [53], StdGen [17], and MAR-3D [6] enable efficient synthesis, while GaussianCube [65], HGMM Splatting [61], and HyperSDFusion [29] improve quality via structured latent spaces. For part-aware generation, SPAGHETTI [18], AutoPartGen [8], and PartGen [7] introduce part-level modeling. However, these methods still struggle with accurate condition-to-3D alignment, leading to semantic inconsistencies.

Schr''odinger Bridge for 3D Generation. Schrödinger Bridge (SB) formulates distribution transport as optimal control, enabling data-to-data generation, rooted in diffusion bridges [11, 9] and classical SB theory [47, 30]. It has been applied to image translation [31, 37], protein matching [50], point cloud denoising [54], and text-to-speech [10]. In 3D, TraCe [33] targets text-to-3D, P2P-Bridge [54] addresses point cloud denoising, and BridgeShape [26] uses latent diffusion bridges for shape completion. However, existing SB-based 3D methods assume a shared metric space, ignoring intrinsic manifold differences, which distorts geometry and degrades fidelity and semantic consistency.

Representation Learning. Neural representation similarity has been widely studied using metrics such as CCA and its variants [21, 45, 40], as well as nonlinear methods like CKA [27] and RSA [28]. The Platonic Representation Hypothesis [23] suggests convergence of representations across modalities with scale, implying universal structure in embeddings. The Aristotelian Representation Hypothesis [16] further emphasizes preservation of local topology across spaces. This motivates cross-modal topological consistency, where local structural invariants remain stable during transport between aligned manifolds.

3 Preliminaries

3D Gaussian Splatting. We adopt 3D Gaussian Splatting (3DGS) [25] as our 3D representation due to its high rendering quality and real-time efficiency. 3DGS represents scenes as a set of anisotropic 3D Gaussians with learnable parameters. Each Gaussian primitive $\mathcal{G}_i$ is parameterized by a center position $\mathbf{x}_i \in \mathbb{R}^3$, scaling factors $\mathbf{s}_i \in \mathbb{R}^3$, rotation quaternion $\mathbf{q}_i \in \mathbb{R}^4$, opacity $\alpha_i \in \mathbb{R}$, and color features $\mathbf{c}_i \in \mathbb{R}^3$. During rendering, Gaussians are projected into 2D screen space and composited via depth-sorted α-blending.

Schrödinger Bridge. The Schrödinger Bridge (SB) problem [47, 30] formulates optimal transport between two probability distributions as an entropy-regularized dynamical problem. Given *paired* data $(\mathbf{x}_0, \mathbf{x}_1)$ where $\mathbf{x}_0 \sim p_{\text{data}}$ and $\mathbf{x}_1 \sim p_{\text{prior}}$, the SB with Doob h-transform [14, 66] defines the forward process as:

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\, dt + g(t)^2\, \mathbf{h}(\mathbf{x}_t, t, \mathbf{x}_1, T)\, dt + g(t)\, d\mathbf{w}_t, \quad \mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x}), \quad \mathbf{x}_T = \mathbf{x}_1 \tag{1}$$

which transports data samples toward the endpoint $\mathbf{x}_1$ at time T. The corresponding backward process used for generation is:

$$d\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t)\, (\mathbf{s}(\mathbf{x}_t, t, \mathbf{x}_1, T) - \mathbf{h}(\mathbf{x}_t, t, \mathbf{x}_1, T))\right] dt + g(t)\, d\bar{\mathbf{w}}_t,\ \mathbf{x}_T = \mathbf{x}_1,\ \mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x}) \tag{2}$$

where $\mathbf{h}(\mathbf{x}, t, \mathbf{x}_1, T) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_T \mid \mathbf{x}_t)\big|_{\mathbf{x}_t=\mathbf{x},\, \mathbf{x}_T=\mathbf{x}_1}$ is the gradient of the log transition kernel of the reference SDE from time t to T, evaluated at $\mathbf{x}_t = \mathbf{x}$ and $\mathbf{x}_T = \mathbf{x}_1$; and $\mathbf{s}(\mathbf{x}, t, \mathbf{x}_1, T) =$

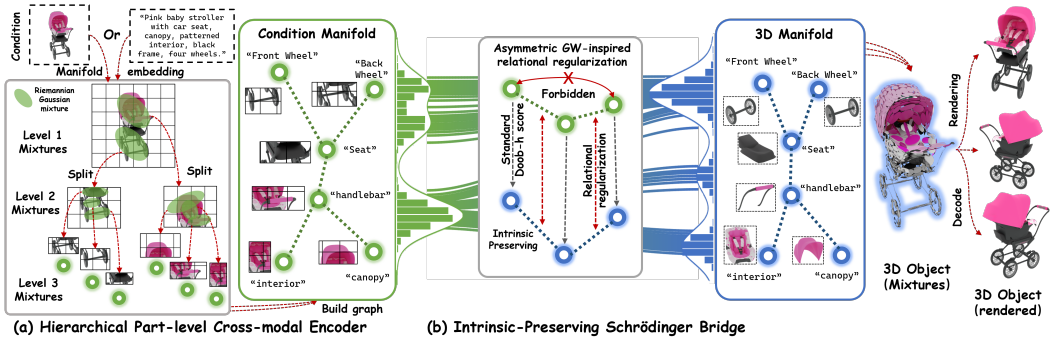

Figure 2: Overview of our method. (a) Hierarchical Part-level Cross-modal Encoder builds hierarchical part-level representations for given conditions and build the part graph for transportation. (b) Intrinsic-Preserving Schrödinger Bridge leverage asymmetric GW-inspired relational regularization to transport from the part graph of condition manifold to 3D manifold while preserving intrinsic structure, enabling direct part-aware 3D generation.

$\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t \mid \mathbf{x}_T)\big|_{\mathbf{x}_t=\mathbf{x},\, \mathbf{x}_T=\mathbf{x}_1}$ is the score of the marginal conditioned on the endpoint. We provide in the Appendix the analytical forms of $\mathbf{x}_t$ at each timestep t under Gaussian kernels.

4 Method

We aim to generate part-aware 3DGS primitives from multimodal conditions with compositional semantics. Our key insight is that multimodal feature spaces share a common neighborhood topology, enabling intrinsic-preserving optimal transport for semantically consistent part generation. As shown in Fig. 2, we propose a hierarchical part-level cross-modal encoder that organizes multi-granularity features of each modality into part graph to provide intrinsic priors (Sec. 4.1). We then introduce Intrinsic-Preserving Schrödinger Bridge with asymmetric GW-inspired relational regularization for intrinsic-preserving cross-modal transport (Sec. 4.2). Finally, we design Compositionally Controllable 3D Generation that fixes unchanged transport paths while independently editing or resampling parts without full regeneration (Sec. 4.3).

4.1 Hierarchical Part-level Cross-modal Encoder

Hierarchical Riemannian Gaussian mixture. Given multimodal data including 3DGS $\mathcal{G}$, images $\mathcal{I}$, and text embeddings $\mathcal{T}$, we first define each mixture layer on the modality-specific manifold $\mathcal{M}$. As each modality resides on a distinct manifold with its own metric, we define a modality-specific metric $d_{\mathcal{M}}$ to better capture its intrinsic geometry. For 3DGS, the $d_{\mathcal{M}}$ is a product-space distance computed over each component with its corresponding metric: $\mathbb{R}^3 \times SO(3) \times \mathbb{R}^+ \times \mathbb{R}^c$ for position, rotation, scale, and color, for images is $\mathbb{R}^2 \times \mathbb{R}^3$, and text lie on the embedding $\mathbb{R}^d$ with cosine distance. Based on $d_{\mathcal{M}}$, we generalize the Gaussian density function to a Riemannian Gaussian density defined on the manifold of each modality:

$$f\left(\Theta_i|\boldsymbol{\theta}_j\right) = \frac{1}{Z} \exp\left(-\frac{d^2_{\mathcal{M}}(\Theta_i, \boldsymbol{\theta}_j)}{2\sigma_j^2}\right), \quad \boldsymbol{\theta}_j = \mathcal{D}(\mathbf{z}_j) \tag{3}$$

where Z is the normalizing factor depend on $d_{\mathcal{M}}$, $\Theta \in \{\mathcal{G}, \mathcal{I}, \mathcal{T}\}$ denotes the input data from any modality, $\boldsymbol{\theta}_j$ represents the parameters decoded from the semantic latent $\mathbf{z}_j$ via parameter decoder $\mathcal{D}$, σ_j is the bandwidth. Given the Riemannian density function, we define each mixture layer as:

$$p\left(\Theta|\Omega^l\right) = \prod_{i=1}^{N} \sum_{j=1}^{J} \pi_j f\left(\Theta_i|\boldsymbol{\theta}_j^l\right), \tag{4}$$

where Ω^l and $\boldsymbol{\theta}^l$ is the mixture model and parameters at level l respectively, π_j is the mixture weight. To maintain the consistency of local and global semantics, we define hierarchical splitting to generate the $\mathbf{z}^{l+1}$ of the next layer mixture model from the current layer $\mathbf{z}^l$. We apply a cross-attention

mechanism to update the semantics of each child latent from the same parent latent, followed by an MLP split $\mathbb{M}$:

$$\mathbf{z}^{l+1} = \sum_{i=1}^{I} \sum_{j=1}^{J} S(j) \mathcal{A}^{i,j} \mathbb{M}^i \left(\mathbf{z}_i^l \right), \tag{5}$$

where $\mathcal{A}^{i,j}$ represents the attention split weights between semantic layers $\{l, l+1\}$, and $S(j)$ denotes the signaling function to ensure that splitting is constrained within its respective child nodes: for $\mathbf{z}_j^{l+1} \in \mathbb{M}(\mathbf{z}^{l+1})$, $S(j) = 1$, in other cases, $S(j) = 0$. We then decode the corresponding parameters $\boldsymbol{\theta}^{l+1}$ via $\mathcal{D}$ to build the next layer mixture model:

$$p\left(\Theta|\gamma^l, \Omega^{l+1}\right) = \prod_{i=1}^{M} \sum_{j=1}^{\hat{J}} \pi_j^{l+1} p\left(\Theta_i | \mathcal{D}(\mathbf{z}_j^{l+1})\right), \quad \gamma_{ij}^l \stackrel{\text{def}}{=} \frac{\pi_j p\left(\Theta_i | \Omega_j^l\right)}{\sum_{j'=1}^{J} \pi_{j'} p\left(\Theta_i | \Omega_{j'}^l\right)}. \tag{6}$$

where γ_{ij}^l serves as a probabilistic mixture assignment propagated across levels.

Part graph construction. From Aristotelian Hypotheses [16], which imply that once mixtures converge the resulting part graphs become comparable, we construct the finest-level part graph ($l = d$) by computing cosine similarities between semantic latents $\mathbf{z}^{l=d}$ and thresholding them to form k-NN edge $\mathcal{E}$:

$$\mathcal{E} = \{(i, j) \mid j \in k\text{NN}(\mathbf{z}_i^{l=d})\}, \tag{7}$$

Although only the finest level is used for graph construction, the hierarchical fitting process is essential to ensure layer-wise consistency and reasonable semantic assignment, preventing redundancy or omission that would result from directly fitting the finest granularity. Finally, a shared decoder $\mathcal{D}_{\text{gs}}$ maps the finest semantic latents $\mathbf{z}^{l=d}$ to 3D Gaussian primitives:$\Theta_j = \mathcal{D}_{\text{gs}}(\mathbf{z}_j^{l=d})$. Subsequently, according to Eq. 4, we apply non-maximum suppression (NMS) to the decoded 3DGS to reconstruct the final 3DGS. Our encoder is trained with a joint optimization objective combining hierarchical likelihood, reconstruction, and rendering losses. We optimize the hierarchical likelihood via:

$$\mathcal{L}_{\text{like}}(\Theta, \Omega) = -\frac{1}{|\Theta|} \sum_{d=1}^{D} \left(l_{\log}\left(\Theta|\Omega^{l=d}\right) + \frac{1}{\sigma^2} \|\mathbf{z}^{l=d}\| \right), \tag{8}$$

when $d = 1$, $l_{\log}(\cdot) = \log p\left(\Theta|\Omega^{l=1}\right)$, else $l_{\log}(\cdot) = \log p\left(\Theta|\gamma^{l=d-1}, \Omega^{l=d}\right)$. We then train the $\mathcal{D}_{\text{gs}}$ via the 3DGS reconstruction and rendering loss $\mathcal{L}_{\text{rr}}$:

$$\mathcal{L}_{\text{rr}} = \|\Theta_p(t) - \Theta_{\text{gt}}(t)\|^2 + \|I_p(t) - I_{\text{gt}}(t)\|^2 + \|\Phi(I_p(t)) - \Phi(I_{\text{gt}}(t))\|^2 \tag{9}$$

where I is rendered image, Φ denotes the VGG feature extractor. By hierarchically constructing part graph within each modality manifold, we obtain representations that share approximately isomorphic neighborhood structures across each modality. Same as GW-OT [42], our Intrinsic-Preserving SB required for GW matching of intrinsic-preserving during transportation in Sec. 4.2.Compared with HGMM Splatting [61], which models 3DGS hierarchies to condition cascaded diffusion, we extend hierarchical encoding to multimodal manifolds with modality-specific metrics and Riemannian mixture densities. Our encoder introduces a simpler intrinsic consistency prior, substantially reducing the difficulty of intrinsic-preserving mapping while improving the consistency and quality of cross-modal transport.

4.2 Intrinsic-Preserving Schrödinger Bridge

Having established the part graph with explicit topological edges in Sec. 4.1, we now consider intrinsic-preserving transport. Consider a standard GW [42] cost:

$$\min_{\pi \in \Pi(p_0, p_1)} \iint (d_0^{i,j} - d_1^{k,l})^2 \, d\pi(i,k) \, d\pi(j,l) \tag{10}$$

where $\pi(\cdot)$ denotes the transport plan between source and target parts, and $d_0^{i,j}$, $d_1^{k,l}$ are intrinsic metrics within their respective manifolds under marginals p_0 and p_1. The standard GW cost simultaneously (1) drives each source part toward its corresponding target and (2) preserves the intrinsic topology of the source manifold along the transport trajectory. Given the part graphs in Sec. 4.1, both marginals p_0 (condition) and p_1 (3D) are available, establishing global distribution-level

alignment. However, fine-grained part correspondences remain ambiguous. Moreover, the Doob h-transformation in Eq. 1 acts independently on nodes without enforcing intrinsic consistency, failing to preserve intrinsic relationships and potentially dispersing semantics during transport.

To address this, we first use the GW cost in Eq. 10 to establish part correspondences between condition and 3D part graphs via entropic-regularized GW with Sinkhorn iterations. Since our part graphs produces approximately isomorphic neighborhood structures, this yields discrete part matches (i, k) efficiently. Given these matched part-level marginals, we apply the simulation-free Doob h-transform directly to each corresponding part pair for point-to-point transport. We simultaneously incorporate the intrinsic-preserving GW objective $d_0^{i,j}$ and $d_1^{k,l}$ by augmenting the h-transform with an asymmetric GW-inspired relational regularization term, enforcing intrinsic-preserving transport across all parts. We thus define the metric d_t^{ij} between nodes i and j as the graph geodesic distance over the part graph:

$$d_t^{ij} = \min_{p: i \rightsquigarrow j} \sum_{(u,v)\in p} w_{uv} \|\mathbf{z}_t^u - \mathbf{z}_t^v\|_2, \tag{11}$$

where t is the timestep of transportation defined in Eq. 1, w_{uv} denotes the edge weight of $(u, v) \in \mathcal{E}$ and $\mathbf{z}_t^u$ is the semantic latent of node u at time t. The initial distance d_0^{ij} and target distance d_1^{ij} are defined analogously on the source and target part graphs, respectively. Building upon the metric and simplified GW objective, we define the asymmetric GW-inspired relational regularization term R_{asym} as:

$$R_{\text{asym}} = \alpha \sum_{(i,j)\in\mathcal{E}_{\text{src}}} w_{ij} \big(d_t^{ij} - d_0^{ij}\big)^2 + \beta \sum_{(i,j)\in\mathcal{E}_{\text{tgt}}} w_{ij} \big(d_t^{ij} - d_1^{ij}\big)^2, \tag{12}$$

where $\mathcal{E}_{\text{src}}$ and $\mathcal{E}_{\text{tgt}}$ denote the edge sets of the source (condition) and target (3D) part graphs, respectively. The coefficients α and β are asymmetric weights with $\alpha > \beta$, reflecting that violations in the source space neighborhood incur high semantic cost while violations in the target space neighborhood incur only low geometric deformation cost. We integrate asymmetric GW-inspired relational regularization into the SB framework by augmenting the h-transform with the regularization gradient. The regularization enhanced h-transform $\mathbf{h}_\theta$ is:

$$\mathbf{h}_\theta(\mathbf{z}_t, t, \mathbf{z}_1, T) = \mathbf{h}(\mathbf{z}_t, t, \mathbf{z}_1, T) + \lambda \nabla_{\mathbf{z}_t} R_{\text{asym}}(\mathbf{z}_t, \mathcal{E}_{\text{src}}, \mathcal{E}_{\text{tgt}}), \tag{13}$$

where $\mathbf{h}(\cdot)$ is the standard Doob h-transform defined in Eq. 1 and λ controls the strength of asymmetric GW-inspired relational regularization. $\nabla_{\mathbf{z}_t} R_{\text{asym}}$ can be obtained analytically via R_{asym}. Since each bridge is conditioned on a fixed target endpoint $\mathbf{z}_1$, the boundary conditions are structurally pinned and unaffected by $\nabla_{\mathbf{z}_t} R_{\text{asym}}$, which is proved in Appendix. The forward and backward SDE formulations remain consistent with the preliminary framework in Eq. 1 and Eq. 2, with only the h-transform being enhanced. The training objective inherits the score matching formulation:

$$\mathcal{L} = \mathbb{E}\big[w(t)\|\mathbf{s}_\theta(\mathbf{z}_t, t, \mathbf{z}_1, T) - \mathbf{h}_\theta(\mathbf{z}_t, t, \mathbf{z}_1, T)\|^2\big], \tag{14}$$

where $\mathbf{s}_\theta$ is the score network used to approximate $\mathbf{h}_\theta$ and $w(t)$ is the time weighting function. After IPSB transport, the resulting $\mathbf{z}$ is decoded by $\mathcal{D}_{\text{gs}}$ into 3DGS primitives, which are then used to reconstruct the complete 3D representation via NMS, following Sec. 4.1. By incorporating asymmetric GW-inspired relational regularization, our IPSB preserves semantic neighborhoods while completing cross-modal generation, establishing the foundation for the part-level intervention mechanism presented in Sec. 4.3.

4.3 Compositionally Controllable 3D Generation

Building upon the IPSB in Sec. 4.2, our intervention strategy partitions the part graph nodes into target and non-target sets. Let $\mathcal{V} = \{\mathbf{z}\}$ and $\mathcal{S} \subset \mathcal{V}$ denote the set of target part nodes selected for intervention, with $\bar{\mathcal{S}} = \mathcal{V} \setminus \mathcal{S}$ representing the frozen complement. Specifically, the new condition c' is first encoded by the cross-modal encoder in Sec. 4.1 to obtain $\mathbf{z}^{c'}$. Target nodes are then selected as $\mathcal{S} = \{j \mid \text{sim}(\mathbf{z}_j, \mathbf{z}^{c'}) > \tau\}$, where τ is the same threshold as Eq. 7. For the selected nodes, we support two intervention modes that share a unified inference framework: **(1)** Condition replacement substitutes the original condition c with a new condition c' for all nodes in $\mathcal{S}$ and reruns the SB transport under the updated guidance; **(2)** Resampling preserves the original condition but draws fresh noise from the conditional distribution to explore alternative realizations of the same semantic

Algorithm 1 Part-wise Intervention Inference

Require: trained model s_θ, source condition c, new condition c', part set $\mathcal{S}$, part graph $(\mathcal{V}, \mathcal{E})$

Run forward pass with c to obtain full trajectories $\{\mathbf{z}_j(t)^*\}$ for all $j \in \mathcal{V}$
Identify $\mathcal{S}$ via similarity and fix trajectories: $\mathbf{z}_j(t) \leftarrow \mathbf{z}_j(t)^*$ for all $j \in \bar{\mathcal{S}}$
for $t = T$ **to**0 **do**
    for $j \in \mathcal{S}$ **do**
        Compute boundary gradient $\nabla R^{\mathcal{S}}_{\text{asym}}$ over edges crossing $\mathcal{S}/\bar{\mathcal{S}}$ for $\mathbf{h}_\theta(\mathbf{z}_j, t, \mathbf{z}_j^{c'}, T)$
        $\mathbf{z}_j(t - dt) \leftarrow \text{SDE_step}(\mathbf{z}_j(t), \mathbf{h}_\theta + \lambda \nabla R^{\mathcal{S}}_{\text{asym}})$
    end for
end for
Decode edited 3D Gaussian: $\Theta' = \mathcal{D}_{\text{gs}}(\mathbf{z}_j(1))$ for $j \in \mathcal{V}$

Table 1: Ablation studies on manifold metrics, graph construction threshold, and asymmetric GW.

Metric	Manifold		k for construction			Asymmetric regularization		
	Riemannian	All Euclidean	$k=2$	$k=4$	$k=8$	$\alpha=\beta$	$\alpha>\beta$	$\alpha<\beta$
NLL↓	**0.97**	1.21	–	–	–	–	–	–
FID↓	**10.26**	16.94	14.32	**10.26**	13.85	15.67	**10.26**	14.58
KID↓	**2.31**	4.96	3.12	**2.31**	2.89	3.89	**2.31**	3.45

part. The intervention mechanism is formalized through a partitioned $\mathbf{h}_\theta$ that distinguishes between active and frozen nodes. The modified drift is:

$$\mathbf{h}^{\text{int}}_\theta(\mathbf{z}_j, t) = \begin{cases} \mathbf{h}_\theta(\mathbf{z}_j, t, \mathbf{z}_j^{c'}, T) + \lambda \nabla_{\mathbf{z}_j} R^{\mathcal{S}}_{\text{asym}} & j \in \mathcal{S} \\ \mathbf{z}^j(t)^* & j \in \bar{\mathcal{S}} \end{cases} \tag{15}$$

where $\mathbf{z}_j(t)^*$ denotes the precomputed fixed trajectory from the original forward pass with condition c, and $\mathbf{z}_j^{c'}$ represents the target latent under the new condition c' for the selected nodes. The term $R^{\mathcal{S}}_{\text{asym}}$ is a boundary-restricted regularization that operates exclusively on edges crossing the partition between $\mathcal{S}$ and $\bar{\mathcal{S}}$. Then the boundary regularization is defined as

$$R^{\mathcal{S}}_{\text{asym}} = \alpha \sum_{(i,j)\in\mathcal{E}, i\in\bar{\mathcal{S}}, j\in\mathcal{S}} w_{ij}(d_t^{ij} - d_0^{ij})^2 \tag{16}$$

where only edges with one endpoint in $\bar{\mathcal{S}}$ and the other in $\mathcal{S}$ contribute to the constraint. The sampling process inherits the score matching formulation from Eq. 14, applied exclusively to nodes in $\mathcal{S}$ since nodes in $\bar{\mathcal{S}}$ require no gradient computation. The intervention procedure is summarized in Alg. 1, which first runs a standard forward pass to cache all trajectories, then selectively updates only the target nodes while maintaining boundary consistency through the restricted regularization. This intervention framework requires no retraining and avoids full regeneration of the 3D representation, with semantic boundaries clearly delineated by the part graph structure.

5 Experiments

Datasets and Metrics. We use OmniObject3D [57] and a filtered LVIS subset of Objaverse [12] for text-to-3D and image-to-3D. For 3D-to-3D we use ShapeNet [5] and OmniObject3D, and 3D-FRONT [15] for scene generation . Quality is evaluated with FID [19] and KID [3] over 50K generated and 50K real renderings. Controllability and editing alignment use CLIP-S [44] and human evaluation. All renders are at 512×512. More details are in the Appendix.

Implement Details. We set 3DGS primitives to $64 \times 64 \times 9 = 36{,}864$ per object, matching other GS-based baselines. IPSB uses 50 sampling steps with a high-order sampler [66]. The encoder has 3 levels with $k = 4$. Asymmetric GW-inspired relational regularization uses $\alpha = 1.0$, $\beta = 0.3$, and $\lambda = 0.5$. More details and results are in the Appendix.

5.1 Ablation Studies

Additional ablation studies are provided in the appendix.

Riemannian Gaussian mixture. We ablate the effect of the Riemannian Gaussian mixture by varying manifold metrics and evaluating semantic precision and generation quality. As shown in

Table 2: Comparison of image-to-3D, text-to-3D, and 3D-to-3D generation.

Method	Image-to-3D			Text-to-3D		3D-to-3D	
	LPIPS↓	FID-50K↓	KID-50K↓	CLIP Score↑	Time(s)↓	FID-50K↓	KID-50K↓
G.Cube [65]	0.1768	8.32	2.67	30.56	2.30	21.87	36.35
Photo3D [34]	0.1627	5.73	4.41	32.67	5.64	19.33	27.64
CoPart [13]	0.1597	7.72	2.38	31.69	∼65	18.35	19.37
TraCe [33]	0.1063	5.19	3.97	30.32	∼70	16.34	12.76
DiffGS [4]	0.1098	5.77	2.88	29.91	∼6	16.87	12.35
HGMMS. [61]	0.1326	6.24	2.55	31.24	4.80	15.74	11.86
LGM [53]	0.1715	7.61	3.18	29.84	4.20	22.46	31.72
DiffSplat [35]	0.1218	5.46	2.41	32.18	2.40	14.82	10.91
Ours	**0.1021**	**4.99**	**1.98**	**33.02**	**2.12**	**9.89**	**6.24**

Table 3: Intrinsic-preserving evaluation using calibrated mKNN (c-mKNN), GW distance, Edge preservation, and RSA.

Metric	G.Cube [65]	Photo3D [34]	CoPart [13]	TraCe [33]	DiffGS [4]	Ours w/o GW	**Ours**
c-mKNN ↑	0.052	0.061	0.071	0.059	0.062	0.072	**0.138**
GW distance ↓	0.284	0.317	0.196	0.251	0.273	0.208	**0.113**
Edge pre. (%) ↑	61.7	43.8	72.4	63.7	62.9	68.2	**84.9**
RSA ↑	0.542	0.427	0.681	0.526	0.514	0.503	**0.793**



Figure 3: Visualization of image-to-3D generation. See appendix for more examples.

Tab. 1, Riemannian Gaussian mixtures consistently improve both metrics. This gain comes from manifold-specific metrics that better preserve intrinsic geometry across modalities, reducing semantic distortion and improving part-level representation fidelity. In contrast, Euclidean assumptions impose isotropic distances that blur semantic boundaries and weaken structure modeling.

K in part graph construction. We ablate K-NN neighbors k for the part graph, shown in Tab. 1. Both $k = 2$ and $k = 8$ degrade performance, while $k = 4$ performs best. Small k yields under-connected graphs, weakening GW-inspired regularization and destabilizing intrinsic transport. Large k introduces spurious connections, corrupting structure and misleading regularization. $k = 4$ balances local semantic adjacency and noise avoidance.

Asymmetric GW-inspired relational regularization. We ablate asymmetric design under symmetric ($\alpha = \beta$), source-prioritized ($\alpha > \beta$, default), and target-prioritized ($\alpha < \beta$) settings. As shown in Tab. 1, symmetric gives moderate results but underuses directional cross-modal structure. Target-prioritized hurts performance by over-constraining target deformation, limiting 3D flexibility. Source-prioritized performs best by preserving semantic neighborhoods in the conditional space while allowing sufficient geometric freedom, validating the asymmetric formulation Eq. 12.

5.2 Experimental Protocol

We provide statistical analyses, along with additional evaluations and visualizations in appendix.

Image and Text-to-3D generation. We evaluate our method on image-to-3D and text-to-3D tasks. Quantitative results are in Tab. 2, with visual comparisons in Fig. 3 and Fig. 4. For image-to-3D, our method consistently outperforms all baselines in all metrics, showing stronger detail preservation and geometry fidelity. For text-to-3D, it achieves the highest CLIP score with the fastest inference time, due to the part-graph build via our encoder, enabling precise alignment via IPSB. In contrast,



Figure 4: Visualization of text-to-3D generation. See appendix for more examples.

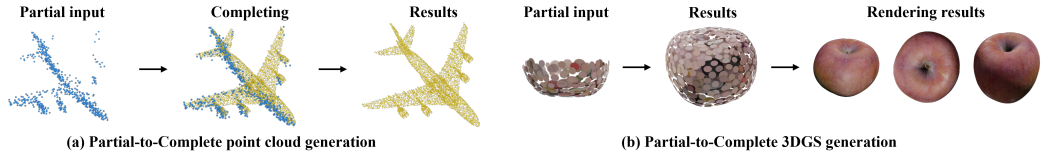

Figure 5: Visualization of 3D-to-3D generation

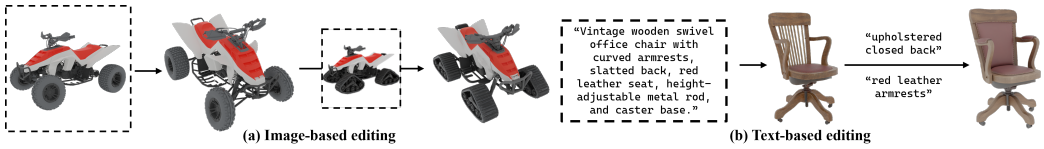

Figure 6: Visualization of image-based and text-based part-aware editing.

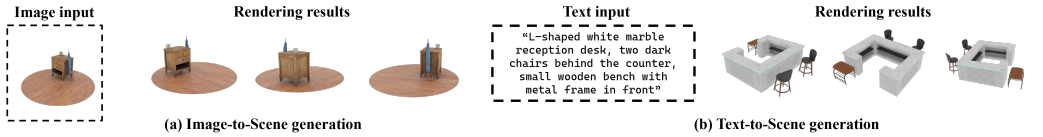

Figure 7: Visualization of image-based and text-based scene generation.

diffusion-based DiffGS and Photo3D are limited by their paradigm, leading to weaker input–output alignment. SB-based TraCe ignores cross-modal metric discrepancies, degrading performance. CoPart improves via part awareness but remains constrained by its diffusion paradigm.

Intrinsic-preserving evaluation. We evaluate intrinsic-preserving consistency using calibrated mKNN (c-mKNN) [16], which measures local neighborhood preservation after removing nearest-neighbor overlap bias via a permutation-based null model. To further evaluate structural preservation, we construct part graphs for the generated outputs and compare them with the conditioning graphs using three additional metrics: Gromov–Wasserstein (GW) distance measures the discrepancy between their pairwise relational structures; edge preservation rate measures the proportion of preserved neighborhood connections under the soft assignment induced by GW alignment; and RSA measures the Spearman correlation between the aligned graph-geodesic distance matrices, assessing consistency in the ordering of pairwise distances. As shown in Tab. 3, our method achieves the highest c-mKNN, indicating best structural alignment between input and output. Removing our GW-inspire regularization significantly degrades results, confirming the role of IPSB. DiffGS and Photo3D show poor consistency due to disrupted local structure from diffusion paradigm. TraCe improves alignment but lacks explicit cross-modal intrinsic constraints. CoPart partially mitigates this via part-aware design, yet remains limited by its diffusion backbone.

3D-to-3D generation. We further evaluate 3D-to-3D generation by retraining baselines with 3D models as priors or conditioning signals. As shown in Tab. 2, our method outperforms all prior work across all metrics. This stems from a hierarchical part-level cross-modal encoder that handles heterogeneous inputs and IPSB, which enables intrinsic-preserving transfer across manifolds, yielding a general-purpose framework. In contrast, prior methods are tailored to text- or image-conditioned settings and lack such generality. We show examples in Fig. 5 , demonstrating strong cross-manifold translation within the same modality, confirming overall generalization ability.

Table 4: Comparison of controllability, part-level editing, and scene generation.

Method	Controllability		Part-level Editing		Scene Generation	
	CLIP-S↑	Human(0-5)↑	CLIP-S↑	Human(0-5)↑	FID-50K↓	KID-50K↓
LGM [53]	26.64	3.7	25.89	3.1	17.54	16.88
G.Cube [65]	24.37	3.1	23.94	2.9	21.33	32.83
DiffGS [4]	27.33	3.8	25.61	3.5	15.78	11.63
CoPart [13]	28.91	4.1	28.97	4.0	15.95	12.37
Photo3D [34]	28.14	3.8	26.37	3.4	15.54	12.08
TraCe [33]	27.62	3.7	25.21	3.2	16.28	13.06
DiffSplat [35]	29.18	4.1	27.73	3.8	15.31	11.86
HGMMS. [61]	29.42	4.3	29.35	4.4	15.48	11.94
Ours	**30.38**	**4.6**	**30.13**	**4.7**	**14.97**	**11.54**

Part-aware controllability and editing. We evaluate part-level controllability under semantic control and part editing. As shown in Tab. 4, our method outperforms baselines in both settings. Fig. 6 shows image- and text-conditioned edits. This gains from an explicit part graph and an intervention mechanism enabling independent semantic manipulation without full regeneration. With IPSB, edits remain structurally optimal compared to diffusion-, SDS-, and feed-forward methods. Additional part-level evaluation are in the Appendix.

Scalability to scene. We explore the potential of our method for scene-level generation. We present image-conditioned and text-conditioned scene generation results in Fig. 7. As presented in Tab. 4, our approach surpasses all baselines across all evaluation metrics. This is because our hierarchical semantic modeling scales adaptively with the input, precisely localizes object-level semantics, and effectively guides the generation process, whereas prior methods do not account for scalability to scene-level inputs.

Complexity analysis. We report total FLOPs in Tab. 5, including the hierarchical part-level cross-modal encoder and IPSB under a fixed batch size. The encoder provides efficient multi-scale semantic encoding with low overhead. Compared to GaussianCube, our method achieves better performance with lower FLOPs. Compared to DiffSplat, it improves results while reducing complexity via hierarchical encoding and efficient SB transport with asymmetric GW regularization. CoPart, despite part awareness, incurs higher cost due to multi-level part latent modeling.

Table 5: Complexity analysis.

Method	GFLOPs↓	FID↓
CoPart [13]	35.33	10.35
DiffSplat [35]	68.46	11.93
G.Cube [65]	25.46	15.67
HGMMS. [61]	31.29	13.27
Photo3D [34]	57.62	11.42
TraCe [33]	86.73	10.71
DiffGS [4]	54.38	10.94
LGM [53]	42.15	14.36
Ours	**25.94**	**10.67**

6 Conclusion

We present the Intrinsic-Preserving Schrödinger Bridge for direct part-aware 3D generation. Our hierarchical part-level cross-modal encoder constructs part graphs across modalities through Riemannian Gaussian mixture modeling and explicit intrinsic topology constraints. Building upon this, our intrinsic-preserving schrödinger bridge employs asymmetric GW-inspired relational regularization to strictly preserve intrinsic structures in the conditional space while allowing sufficient geometric flexibility in the 3D manifold. Furthermore, our part-wise intervention inference enables independent replacement or resampling of target part nodes for compositionally controllable 3D generation.

Limitations. Future work will explore adaptive regularization to improve flexibility and quality.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant (62373293, 62463020, 62472336, 62302369, U2570205); in part by the Science and Technology Project of Jiangxi province (20261BCE310009, 20261BCE310010, 20252BAC200185), in part by Young Science and Technology Star Program of Shaanxi Province (2025ZC-KJXX-25) and in part by Ganpo Talent Support Plan under Grant gpjc20260121.

References

[1] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[2] Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296*, 2025.

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[4] Yuanhao Cai, He Zhang, Kai Zhang, Yixun Liang, Mengwei Ren, Fujun Luan, Qing Liu, Soo Ye Kim, Jianming Zhang, Zhifei Zhang, et al. 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*, pages 25062–25072, 2025.

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[6] Jinnan Chen, Lingting Zhu, Zeyu Hu, Shengju Qian, Yugang Chen, Xin Wang, and Gim Hee Lee. Mar-3d: Progressive masked auto-regressor for high-resolution 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11083–11092, 2025.

[7] Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, and Andrea Vedaldi. Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5881–5892, 2025.

[8] Minghao Chen, Jianyuan Wang, Roman Shapovalov, Tom Monnier, Hyunyoung Jung, Dilin Wang, Rakesh Ranjan, Iro Laina, and Andrea Vedaldi. Autopartgen: Autoregressive 3d part generation and discovery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[9] Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood training of schrödinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*, 2021.

[10] Zehua Chen, Guande He, Kaiwen Zheng, Xu Tan, and Jun Zhu. Schrodinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*, 2023.

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

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.

[13] Shaocong Dong, Lihe Ding, Xiao Chen, Yaokun Li, Yuxin Wang, Yucheng Wang, Qi Wang, Jaehyeok Kim, Chenjian Gao, Zhanpeng Huang, et al. From one to more: Contextual part latents for 3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8230–8240, 2025.

[14] Joseph L Doob and JI Doob. *Classical potential theory and its probabilistic counterpart*, volume 262. Springer, 1984.

[15] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021.

[16] Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

[17] Yuze He, Yanning Zhou, Wang Zhao, Zhongkai Wu, Kaiwen Xiao, Wei Yang, Yong-Jin Liu, and Xiao Han. Stdgen: Semantic-decomposed 3d character generation from single images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26345–26355, 2025.

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

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[21] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in Statistics: Methodology and Distribution*. Springer, 1992.

[22] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

[23] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[26] 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, pages 5726–5734, 2026.

[27] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

[28] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008.

[29] Zhiying Leng, Tolga Birdal, Xiaohui Liang, and Federico Tombari. Hypersdfusion: Bridging hierarchical structures in language and geometry for enhanced 3d text2shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19691–19700, 2024.

[30] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems-A*, 34(4):1533–1574, 2014.

[31] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 2023.

[32] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[33] Ziying Li, Xuequan Lu, Xinkui Zhao, Guanjie Cheng, Shuiguang Deng, and Jianwei Yin. Walking the schr\" odinger bridge: A direct trajectory for text-to-3d generation. *arXiv preprint arXiv:2511.05609*, 2025.

[34] Xinyue Liang, Zhinyuan Ma, Lingchen Sun, Yanjun Guo, and Lei Zhang. Photo3d: Advancing photorealistic 3d generation through structure-aligned detail enhancement. *arXiv preprint arXiv:2512.08535*, 2025.

[35] Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. Diffsplat: Repurposing image diffusion models for scalable gaussian splat generation. *arXiv preprint arXiv:2501.16764*, 2025.

[36] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[37] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I ^{2} sb: Image-to-image schrödinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

[38] Yibo Liu, Zheyuan Yang, Guile Wu, Yuan Ren, Kejian Lin, Bingbing Liu, Yang Liu, and Jinjun Shan. Vqa-diff: Exploiting vqa and diffusion for zero-shot image-to-3d vehicle asset generation in autonomous driving. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024.

[39] Bo Miao, Mingtao Feng, Zijie Wu, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Referring human pose and mask estimation in the wild. *Advances in Neural Information Processing Systems*, 37:44791–44813, 2024.

[40] Ari S Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 2018.

[41] KL Navaneet, Kossar Pourahmadi Meibodi, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Compgs: Smaller and faster gaussian splatting with vector quantization. In *European Conference on Computer Vision*, pages 330–349. Springer, 2024.

[42] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.

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

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[45] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017.

[46] Salem Said, Lionel Bombrun, Yannick Berthoumieu, and Jonathan H Manton. Riemannian gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory*, 63(4):2153–2170, 2017.

[47] Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310, 1932.

[48] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[49] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schr\" odinger bridge matching. *arXiv preprint arXiv:2303.16852*, 2023.

[50] Vignesh Ram Somnath, Matteo Pariset, Ya-Ping Hsieh, Maria Rodriguez Martinez, Andreas Krause, and Charlotte Bunne. Aligned diffusion schr\" odinger bridges. *arXiv preprint arXiv:2302.11419*, 2023.

[51] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

[52] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[53] 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 *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[54] Mathias Vogel, Keisuke Tateno, Marc Pollefeys, Federico Tombari, Marie-Julie Rakotosaona, and Francis Engelmann. P2p-bridge: Diffusion bridges for 3d point cloud denoising. In *European Conference on Computer Vision*, pages 184–201. Springer, 2024.

[55] Yanwen Wang, Yiyu Zhuang, Jiawei Zhang, Li Wang, Yifei Zeng, Xun Cao, Xinxin Zuo, and Hao Zhu. Tera: Rethinking text-guided realistic 3d avatar generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10686–10697, 2025.

[56] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 37:121859–121881, 2024.

[57] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 803–814, 2023.

[58] Zijie Wu and Na Zhao. Hierascaffold: Learning compact hierarchical representations for scalable 4d lidar generation. In *Forty-third International Conference on Machine Learning*.

[59] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8929–8939, 2023.

[60] Hongchi Xia, Chih-Hao Lin, Hao-Yu Hsu, Quentin Leboutet, Katelyn Gao, Michael Paulitsch, Benjamin Ummenhofer, and Shenlong Wang. Holoscene: Simulation-ready interactive 3d worlds from a single video. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[61] Qitong Yang, Mingtao Feng, Zijie Wu, Weisheng Dong, Fangfang Wu, Yaonan Wang, and Ajmal Mian. Hierarchical gaussian mixture model splatting for efficient and part controllable 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11104–11114, 2025.

[62] Qitong Yang, Mingtao Feng, Zijie Wu, Shijie Sun, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Learning coherent matrixized representation in latent space for volumetric 4d generation. *Pattern Recognition*, page 113531, 2026.

[63] Yu Yang, Alan Liang, Jianbiao Mei, Yukai Ma, Yong Liu, and Gim Hee Lee. X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[64] LAN Yushi, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud flow matching for 3d generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[65] Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: A structured and explicit radiance representation for 3d generative modeling. *Advances in Neural Information Processing Systems*, 37:97445–97475, 2024.

[66] Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *The Twelfth International Conference on Learning Representations*, 2024.

[67] Dekai Zhu, Yan Di, Stefan Gavranovic, and Slobodan Ilic. Sealion: Semantic part-aware latent point diffusion models for 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11789–11798, 2025.

A Math Preliminary of Schrödinger Bridge

For SB processes with Gaussian transition kernels, *e.g* VE, VP [51], the sampling distribution is a Gaussian distribution of the form according to DDBM [66]:

$$\begin{aligned} q(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_T) = \mathcal{N}(\hat{\mu}_t, \hat{\sigma}_t^2 \boldsymbol{I}), \quad \text{where} \\ \hat{\mu}_t = \frac{\text{SNR}_T}{\text{SNR}_t}\frac{\alpha_t}{\alpha_T}\mathbf{x}_T + \alpha_t \mathbf{x}_0 (1 - \frac{\text{SNR}_T}{\text{SNR}_t}) \\ \hat{\sigma}_t^2 = \sigma_t^2 (1 - \frac{\text{SNR}_T}{\text{SNR}_t}) \end{aligned} \tag{17}$$

where α_t and σ_t are pre-defined signal and noise schedules and $\text{SNR}_t = \alpha_t^2/\sigma_t^2$ is the signal-to-noise ratio at time t. We present both VP and VE setting in Tab. 6 and recommend choosing VP mode.

Table 6: VP and VE setting of SB process.

	$\mathbf{f}(\mathbf{x}_t, t)$	$g^2(t)$	$p(\mathbf{x}_t \mid \mathbf{x}_0)$	SNR_t	$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_T \mid \mathbf{x}_t)$
VP	$\frac{d \log \alpha_t}{dt}\mathbf{x}_t$	$\frac{d}{dt}\sigma_t^2 - 2\frac{d \log \alpha_t}{dt}\sigma_t^2$	$\mathcal{N}(\alpha_t \mathbf{x}_0, \sigma_t^2 \boldsymbol{I})$	α_t^2/σ_t^2	$\frac{(\alpha_t/\alpha_T)\mathbf{x}_T - \mathbf{x}_t}{\sigma_t^2(\text{SNR}_t/\text{SNR}_T - 1)}$
VE	$\mathbf{0}$	$\frac{d}{dt}\sigma_t^2$	$\mathcal{N}(\mathbf{x}_0, \sigma_t^2 \boldsymbol{I})$	$1/\sigma_t^2$	$\frac{\mathbf{x}_T - \mathbf{x}_t}{\sigma_T^2 - \sigma_t^2}$

B Analysis of Asymmetric GW-inspired Relational Regularization

B.1 Theoretical Analysis of Boundary Conditions

We prove that augmenting the Doob h-transform with the asymmetric GW-inspired relational regularization does not alter the boundary conditions of the IPSB. Recall that the gradient entering the enhanced h-transform:

$$\nabla_{\mathbf{z}_t} R_{\text{asym}} = 2\alpha \sum_{(i,j)\in\mathcal{E}_{\text{src}}} w_{ij}\big(d_t^{ij} - d_0^{ij}\big)\nabla_{\mathbf{z}_t} d_t^{ij} + 2\beta \sum_{(i,j)\in\mathcal{E}_{\text{tgt}}} w_{ij}\big(d_t^{ij} - d_1^{ij}\big)\nabla_{\mathbf{z}_t} d_t^{ij}, \tag{18}$$

where d_t^{ij} is the geodesic distance for part graph.

Proposition 1 (Boundary Preservation)**.** *Let $(\mathbf{z}_t)_{t\in[0,T]}$ be the stochastic process governed by the enhanced h-transform $\mathbf{h}_\theta$. Then the boundary conditions of the IPSB are preserved:*

(i) (Initial condition) $\mathbf{z}_0 \sim p_0$ almost surely.

(ii) (Terminal condition) $\mathbf{z}_T = \mathbf{z}_1$ almost surely.

Proof. **(i) Initial condition.** The regularization gradient $\lambda\nabla_{\mathbf{z}_t} R_{\text{asym}}$ enters only as an additive modification to the SDE drift for $t \in (0, T)$. The initial marginal p_0 is a distributional constraint prescribed before the transport process begins, and it is entirely independent of the functional form of the drift. Hence $\mathbf{z}_0 \sim p_0$ is unaffected.

(ii) Terminal condition. We argue in two steps: (a) the regularization gradient is uniformly bounded, and (b) the standard h-transform drift diverges as $t \to T$, dominating the bounded perturbation.

Step (a): Boundedness of $\nabla_{\mathbf{z}_t} R_{asym}$. The geodesic distance d_t^{ij}

$$d_t^{ij} = \min_{p:\, i \rightsquigarrow j} \sum_{(u,v)\in p} w_{uv} \|\mathbf{z}_t^u - \mathbf{z}_t^v\|_2 \tag{19}$$

is a finite sum of ℓ_2-norms of latent differences. Its gradient $\nabla_{\mathbf{z}_t} d_t^{ij}$ is therefore Lipschitz continuous in $\mathbf{z}_t$. Since the edge sets $\mathcal{E}_{\text{src}}$ and $\mathcal{E}_{\text{tgt}}$ are finite, R_{asym} is a sum of smooth, polynomial-type functions of the latents, and for any compact neighbourhood there exists a constant $C < \infty$ such that

$$\big\|\nabla_{\mathbf{z}_t} R_{\text{asym}}(\mathbf{z}_t, \mathcal{E}_{\text{src}}, \mathcal{E}_{\text{tgt}})\big\|_2 \;\le\; C, \qquad \forall\, t \in [0, T]. \tag{20}$$

Step (b): Divergence of the h-transform drift. In the Gaussian bridge framework of DDBM [66], the bridge marginal satisfies $p(\mathbf{z}_t \mid \mathbf{z}_0, \mathbf{z}_T) = \mathcal{N}(\hat{\mu}_t, \hat{\sigma}_t^2 \mathbf{I})$ with

$$\hat{\sigma}_t^2 = \sigma_t^2 \left(1 - \frac{\mathrm{SNR}_T}{\mathrm{SNR}_t}\right) \xrightarrow{t \to T} 0, \tag{21}$$

since $\mathrm{SNR}_t \to \mathrm{SNR}_T$ as $t \to T$. By Bayes' rule the h-transform score can be written as

$$\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_T = \mathbf{z}_1 \mid \mathbf{z}_t) = \frac{\hat{\mu}_t(\mathbf{z}_1) - \mathbf{z}_t}{\hat{\sigma}_t^2}, \tag{22}$$

where $\hat{\mu}_t(\mathbf{z}_1)$ is the posterior mean towards $\mathbf{z}_1$. As $t \to T$, $\hat{\sigma}_t^2 \to 0$ while $\hat{\mu}_t(\mathbf{z}_1) \to \mathbf{z}_1$, so:

$$\left\| g^2(t)\, \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_T = \mathbf{z}_1 \mid \mathbf{z}_t) \right\|_2 \;=\; O\big(\hat{\sigma}_t^{-2}\big) \;\xrightarrow{t \to T}\; +\infty \tag{23}$$

whenever $\mathbf{z}_t \neq \mathbf{z}_1$.

Conclusion. Comparing Eq. 20 and Eq. 23, the regularization perturbation satisfies

$$\left\| \lambda \nabla_{\mathbf{z}_t} R_{\text{asym}} \right\|_2 \;=\; O(1) \;\ll\; \left\| g^2(t)\, \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_T = \mathbf{z}_1 \mid \mathbf{z}_t) \right\|_2, \quad t \to T. \tag{24}$$

By the Girsanov theorem, the law of the enhanced SDE is absolutely continuous with respect to the standard bridge measure on every compact interval $[0, T - \varepsilon]$ for $\varepsilon > 0$. Taking $\varepsilon \to 0$, the singular h-transform drift dominates and forces $\mathbf{z}_t \to \mathbf{z}_1$ as $t \to T$, exactly as in the unperturbed bridge. Therefore $\mathbf{z}_T = \mathbf{z}_1$ almost surely. □

At the terminal time $t = T$ one has $\mathbf{z}_T^u = \mathbf{z}_1^u$ for all nodes u, so $d_T^{ij} = d_1^{ij}$ and the second sum in R_{asym} vanishes identically: $\sum_{(i,j) \in \mathcal{E}_{\text{tgt}}} w_{ij} (d_T^{ij} - d_1^{ij})^2 = 0$. This provides an explicit confirmation that the regularization introduces no residual forcing at the terminal time, and the endpoint is precisely pinned by the diverging h-transform drift.

B.2 Theoretical Analysis of Convergence

We establish two complementary results: (1) the IPSB training objective is a *consistent* estimator of the oracle enhanced h-transform, and (2) the distribution of generated samples converges weakly to the target p_1. We first prove the consistency of our matching objective and the score matching objective.

Theorem B.1 (Score Matching Consistency for IPSB)**.** *The unconstrained minimiser of the IPSB training objective $\mathcal{L}$ satisfies*

$$\mathbf{s}^*(\mathbf{z}_t, t, \mathbf{z}_1, T) \;=\; \mathbf{h}_\theta^*(\mathbf{z}_t, t, \mathbf{z}_1, T) \tag{25}$$

$q(\mathbf{z}_t, \mathbf{z}_1, t)$-almost everywhere, where $\mathbf{h}_\theta^*$ denotes the oracle enhanced h-transform.

Proof. Writing the objective explicitly,

$$\begin{aligned} \mathcal{L} = \int_t \int_{\mathbf{z}_t, \mathbf{z}_1} & q(\mathbf{z}_t, \mathbf{z}_1)\, p(t)\, w(t) \left\| \mathbf{s}_\theta(\mathbf{z}_t, t, \mathbf{z}_1, T) - \mathbf{h}_\theta^*(\mathbf{z}_t, t, \mathbf{z}_1, T) \right\|^2 d\mathbf{z}_t\, d\mathbf{z}_1\, dt \\ & + \; \mathbb{E}\big[w(t) \|\mathbf{h}_\theta^*\|^2\big] \;-\; \mathbb{E}\big[w(t) \|\mathbf{h}_\theta^*\|^2\big]. \end{aligned} \tag{26}$$

Since $p(t) > 0$ and $w(t) > 0$ for $t \in [0, T]$, the functional $\mathbf{s}_\theta \mapsto \mathcal{L}$ is an L^2 loss weighted by the positive measure $w(t)\, p(t)\, q(\mathbf{z}_t, \mathbf{z}_1)$. Its unique point-wise minimiser is

$$\mathbf{s}^*(\mathbf{z}_t, t, \mathbf{z}_1, T) = \mathbb{E}\big[\mathbf{h}_\theta^*(\mathbf{z}_t, t, \mathbf{z}_1, T) \mid \mathbf{z}_t, t, \mathbf{z}_1\big] = \mathbf{h}_\theta^*(\mathbf{z}_t, t, \mathbf{z}_1, T), \tag{27}$$

where the last equality holds because $\mathbf{h}_\theta^*$ is a deterministic function of $(\mathbf{z}_t, t, \mathbf{z}_1, T)$: once $\mathbf{z}_t$, t, and $\mathbf{z}_1$ are conditioned upon, $\mathbf{h}_\theta^*$ is fully determined by $\mathbf{h}_\theta$, d_t^{ij}, and R_{asym}. □

Theorem B.1 parallels Proposition 2 in DDBM [66]. The key observation is that adding the regularization gradient $\lambda \nabla_{\mathbf{z}_t} R_{\text{asym}}$ to the target function does *not* destroy the L^2 structure of the loss: the enhanced target $\mathbf{h}_\theta^*$ is still a deterministic function of the conditioning variables, so the score network $\mathbf{s}_\theta$ can in principle recover it exactly. We then show the weak convergence property of IPSB to the target distribution

Theorem B.2 (Distributional Convergence of IPSB)**.** *Suppose the score network satisfies* $\|\mathbf{s}_\theta - \mathbf{h}_\theta^*\|_{L^2(q(\mathbf{z}_t,\mathbf{z}_1,t))} \le \varepsilon$*. Let* $\hat{p}_1^\theta$ *denote the distribution of the terminal sample* $\mathbf{z}_T$ *generated by running the SDE with* $\mathbf{s}_\theta$*, and let* p_1 *be the target distribution. Then*

$$\mathrm{KL}\big(\hat{p}_1^\theta \,\big\|\, p_1\big) \;\le\; \frac{\lambda_{\max}}{2}\int_0^T w^{-1}(t)\,\varepsilon^2\,dt, \tag{28}$$

where $\lambda_{\max}$ *is a constant depending on the SDE coefficients* $g(t)$ *and the diffusion structure. In particular,* $\hat{p}_1^\theta \to p_1$ *weakly as* $\varepsilon \to 0$*.*

Proof. We decompose the argument into three parts.

Part 1: Correctness of the oracle process. By Proposition 1(ii), the oracle process (using $\mathbf{h}_\theta^*$ exactly) satisfies $\mathbf{z}_T = \mathbf{z}_1$ almost surely, where $\mathbf{z}_1 \sim p_1$ by the bridge construction. Hence the terminal distribution of the oracle process is exactly p_1.

Part 2: Effect of score approximation. Let Q^* and Q^θ denote the path measures induced by the oracle enhanced SDE and the approximated SDE, respectively. Writing the drift discrepancy as $\Delta_t = \mathbf{s}_\theta(\mathbf{z}_t, t, \mathbf{z}_1, T) - \mathbf{h}_\theta^*(\mathbf{z}_t, t, \mathbf{z}_1, T)$, the Girsanov theorem gives

$$\mathrm{KL}(Q^\theta \| Q^*) = \frac{1}{2}\mathbb{E}_{Q^\theta}\left[\int_0^T g^{-2}(t)\|\Delta_t\|^2\,dt\right]. \tag{29}$$

Since $g^{-2}(t)$ and $w(t)$ satisfy $g^{-2}(t) \le \lambda_{\max} w^{-1}(t)$ for a structural constant $\lambda_{\max}$, and $\mathbb{E}[\|\Delta_t\|^2] \le \varepsilon^2$ by assumption, we obtain

$$\mathrm{KL}(Q^\theta \| Q^*) \;\le\; \frac{\lambda_{\max}}{2}\int_0^T w^{-1}(t)\,\varepsilon^2\,dt. \tag{30}$$

Part 3: Marginal KL bound. By the data-processing inequality, the KL divergence between the terminal marginals satisfies

$$\mathrm{KL}\big(\hat{p}_1^\theta \| p_1\big) \;\le\; \mathrm{KL}(Q^\theta \| Q^*) \;\le\; \frac{\lambda_{\max}}{2}\int_0^T w^{-1}(t)\,\varepsilon^2\,dt. \tag{31}$$

As $\varepsilon \to 0$, $\mathrm{KL}(\hat{p}_1^\theta \| p_1) \to 0$, which by Pinsker's inequality implies $\hat{p}_1^\theta \to p_1$ in total variation, and hence weakly. □

The bound in Theorem B.2 is independent of the regularization weights α, β, and λ. This confirms that the IPSB objective is well-posed: increasing the regularization strength does not degrade the asymptotic convergence guarantee; it only changes the intermediate transport trajectory while leaving both boundary conditions and the terminal distribution intact.

B.3 Discussion on GW-inspired regularization

Implicit attention-based models implicitly learn part relations through feature interactions from large-scale data without explicit structural constraints, making relations hard to interpret or control. Our GW-based constraint explicitly models pairwise structures during transport, preserving structure with theoretical guarantees.

Moreover, GW-based constraint also benefits editing and data efficiency. 1) Explicit part graphs enable training-free localized editing (Alg. 1) by freezing selected nodes and optimizing only cross-boundary relations, unlike attention-based methods that require full regeneration. 2) Structural consistency from GW regularization allows strong structure preservation with only 161K objects, whereas attention-based methods typically need 800K objects. Tab. 3 shows this advantage: CoPart achieves c-mKNN 0.071 vs. 0.138 for our method.

C Sampling Algorithm for Intrinsic-Preserving Schrödinger Bridge

Our sampling process is built upon prior higher-order ODE sampler in [24], which discretizes the sampling steps into $t_N > t_{N-1} > \cdots > t_0$ with decreasing intervals, and hybrid Euler-Maruyama

Algorithm 2 IPSB Sampling Algorithm

Input: model s_θ, time steps $\{t_i\}_{i=0}^N$, max time T, guidance strength w, step ratio s
Output: $\mathbf{x}_0$
Sample $\mathbf{x}_N \sim q_{\text{data}}(\mathbf{y})$
for $i = N, \dots, 1$ **do**
  Sample $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\hat{t}_i \leftarrow t_i + s(t_{i-1} - t_i)$
  $\mathbf{d}_i \leftarrow -\mathbf{f}(\mathbf{x}_i, t_i) + g^2(t_i)\left(s_\theta(\mathbf{x}_i, t_i, \mathbf{x}_N, T) - \mathbf{h}_\theta(\mathbf{x}_i, t_i, \mathbf{x}_N, T)\right)$
  $\hat{\mathbf{x}}_i \leftarrow \mathbf{x}_i + \mathbf{d}_i(\hat{t}_i - t_i) + g(t_i)\sqrt{\hat{t}_i - t_i}\boldsymbol{\epsilon}_i$
  $\hat{\mathbf{d}}_i \leftarrow -\mathbf{f}(\hat{\mathbf{x}}_i, \hat{t}_i) + g^2(\hat{t}_i)\left(\frac{1}{2}s_\theta(\hat{\mathbf{x}}_i, \hat{t}_i, \mathbf{x}_N, T) - w\mathbf{h}_\theta(\hat{\mathbf{x}}_i, \hat{t}_i, \mathbf{x}_N, T)\right)$
  $\mathbf{x}_{i-1} \leftarrow \hat{\mathbf{x}}_i + \hat{\mathbf{d}}_i(t_{i-1} - \hat{t}_i)$
  if $i \neq 1$ **then**
    $\mathbf{d}'_i \leftarrow -\mathbf{f}(\mathbf{x}_{i-1}, t_{i-1}) + g^2(t_{i-1})\left(\frac{1}{2}s_\theta(\mathbf{x}_{i-1}, t_{i-1}, \mathbf{x}_N, T) - w\mathbf{h}_\theta(\mathbf{x}_{i-1}, t_{i-1}, \mathbf{x}_N, T)\right)$
    $\mathbf{x}_{i-1} \leftarrow \hat{\mathbf{x}}_i + \left(\frac{1}{2}\mathbf{d}'_i + \frac{1}{2}\hat{\mathbf{d}}_i\right)(t_{i-1} - \hat{t}_i)$
  end if
end for

Table 7: Details of each dataset.

Dataset	#Objects	#Views/object	Rotation	Elevation
ShapeNet	78,653	150	$[0, 2\pi]$	$[\frac{1}{6}\pi, \frac{1}{2}\pi]$
OmniObject3D	5,795	100	$[0, 2\pi]$	$[0, \frac{1}{2}\pi]$
Objaverse	82,575	150	$[0, 2\pi]$	$[0, \frac{2}{3}\pi]$
3D-FRONT	18,968	150	$[0, 2\pi]$	$[0, \frac{2}{3}\pi]$

step in [66] which follows the backward SDE in between higher-order ODE steps to reduce the uncertainty of the marginal distributions at both ends. We further use our score net s_θ and asymmetric GW-inspired relational regularization enhanced score $\mathbf{h}_\theta$ to guide the sampling process, as shown in Algo. 2. We follow the VP setting of guidance strength w and step ratio s in DDBM [66].

D Detailed Experiments Setting

D.1 Implement details.

We use 8 RTX4090 GPUs for training and sampling. We employ 0-degree spherical harmonics to obtain Gaussian primitives. This approach significantly reduces the dimensionality of the Gaussian primitives while preserving fitting quality, thereby simplifying the complexity of diffusion modeling. For other optimization settings, we keep them consistent with the original 3D Gaussian Splatting.

Text inputs use CLIP token embeddings projected to 16-dim via a lightweight 3-layer MLP. Image inputs normalize RGB values and pixel coordinates to $[-1, 1]$. 3D inputs are obtained by fitting 3DGS on Objaverse and ShapeNet objects, then normalizing 3DGS parameters to $[-1, 1]$. The encoder design is shared across modalities. We initialize random root latents and hierarchically enhance them with attention for parent-child information exchange, followed by splitting via a 3-layer MLP that expands latent dimensions and evenly partitions outputs among children. The decoder $\mathcal{D}$ at each level is a 3-layer MLP predicting mixture means, Gram–Schmidt orthonormal bases, eigenvalues, and mixture weights. These parameters construct covariance to implement hierarchical mixture assignments and compute hierarchical log-likelihoods. The 3DGS decoder $\mathcal{D}_{\text{gs}}$ is also a 3-layer MLP predicting 14-dim parameters for each 3DGS primitive. The score/bridge model is a 48-layer DiT-style Graph Transformer with Graphormer attention. Timestep embeddings are injected into each block via DiT's AdaLN-Zero conditioning. Details are provided in Q8.

In our Riemannian Gaussian mixture, we set the number of mixture at final level to 128, and the normalizing factor Z depends on the modality-specific manifold metric $d_{\mathcal{M}}$ and in general admits no closed-form expression for the heterogeneous product spaces considered in this work (*e.g.*, $\mathbb{R}^3 \times SO(3) \times \mathbb{R}^+ \times \mathbb{R}^c$ for 3DGS). In practice, we do not evaluate Z explicitly. This is justified

Table 8: Detailed configuration of generation.

Parameters	ShapeNet	OmniObject3D	Objaverse	3D-FRONT
Training steps	1,000	1,000	1,000	1,000
Inference steps	50	50	50	50
Inference sampler	DDBM-solver	DDBM-solver	DDBM-solver	DDBM-solver

on two grounds. First, Z is shared across all mixture components within the same modality at a given level; it therefore cancels exactly when computing the normalized mixture responsibilities γ_{ij}^{l} via softmax-style normalization. Second, in the hierarchical likelihood objective $\mathcal{L}_{\text{like}}$, omitting a constant additive term $\log Z$ does not change the gradient with respect to the model parameters, leaving optimization unaffected. This approximation is standard in Riemannian mixture models [46] and introduces no systematic bias in our setting.

For each mixture component j at level l, we estimate the bandwidth σ_j adaptively using the decoded component center $\boldsymbol{\theta}_j = \mathcal{D}(\mathbf{z}_j)$ and the k-nearest primitive neighbors under $d_{\mathcal{M}}$:

$$\sigma_j^2 \;=\; \frac{1}{k} \sum_{i \in k\text{NN}(\boldsymbol{\theta}_j)} d_{\mathcal{M}}^2(\Theta_i,\, \boldsymbol{\theta}_j), \tag{32}$$

where Θ_i denotes the i-th input primitive of the corresponding modality. This k-NN-based heuristic provides a local density estimate without requiring a global bandwidth hyperparameter, and it adapts naturally to the varying density of each modality manifold. The encoder is trained with $k = 4$ neighbors consistent with the part graph construction. For scene-level inputs with higher primitive counts, the per-component σ_j adjusts automatically as the local neighborhood density changes.

We report our configuration for generation in Tab. 8, including noise schedule, diffusion steps, and inference sampler.

During inference, given text or image input, we construct its part graph with the trained encoder, feed it into the score network to estimate the score function, and generate samples by iteratively applying the IPSB Sampling Algorithm. For editing, we cache trajectories from the original generation process. By comparing cached trajectories with the new condition, we identify nodes requiring modification and regenerate only corresponding parts via Part-wise Intervention Inference, while preserving trajectories of unchanged parts.

Details of training data. We provide the total number of objects used for training, the number of views rendered per object for obtaining Gaussian primitives, and the distribution of camera poses used for rendering in Tab. 7. We use the same filter from LGM [53] to get the subset of Objaverse dataset. We use the render script to render ShapeNet dataset.

D.2 User study setting.

To further validate the performance of editing and controllability, we defined a controllability metric through a user study. Specifically, users were asked to select the most faithful completion results from multiple methods based on overall quality, different scales, and individual components. We also asked participants to rate specific evaluation criteria on a scale from 0 to 5.

D.3 Discussion of scalability

We argue that mathematical formulation does not hinder the scalability. And the computation of GW distance and the construction of hierarchical GMMs can scale up efficiently, as shwon in Tab. 9. Because 1) Our hierarchical GMMs are designed to simplify complex data. As 3DGS primitive count increases, they provide a more compact abstraction, improving processing efficiency rather than increasing complexity. 2) Most mathematically intensive operations are performed on the part graph rather than the original 36,864 Gaussian primitives. With only 128 default part latents, the compact and adjustable graph representation remains much smaller than the 3DGS set, enabling formulations to scale efficiently to larger primitive counts. 3) We report wall-clock time and peak memory while varying N and J independently in Tab. 9. Increasing N with fixed J leads to near-linear fitting time growth, while Sinkhorn GW and GW regularization costs remain nearly constant. Increasing J raises GW cost as expected, but it remains substantially lower than score prediction cost.

Table 9: Computational scaling of the structural components.

Setting	N	J	HGMM fitting (ms/object)	Sinkhorn-GW (ms/step)	GW regulation (ms/step)	Peak memory (GB)
Primitive scale	9,216	128	5.6	4.8	1.5	8.1
Primitive scale	18,432	128	10.4	4.9	1.5	8.3
Default	**36,864**	**128**	**19.8**	**5.0**	**1.6**	**8.7**
Primitive scale	73,728	128	38.7	5.1	1.6	9.4
Part scale	36,864	32	6.3	0.5	0.2	7.6
Part scale	36,864	64	11.2	1.4	0.5	7.9
Default	**36,864**	**128**	**19.8**	**5.0**	**1.6**	**8.7**
Part scale	36,864	256	37.9	18.6	5.9	12.8

Table 10: Ablation on the dimension of root latent $\mathbf{z}$. Numbers in parentheses indicate variance across 100 repeated fits on 1,000 objects.

Dimension	NLL↓	IoU↑
256	6.23 ($\pm$ 0.019)	0.61 ($\pm$ 0.014)
512	1.02 ($\pm$ 0.017)	0.83 ($\pm$ 0.016)
768	**0.97** ($\pm$ 0.014)	**0.96** ($\pm$ 0.011)
896	0.96 ($\pm$ 0.013)	0.96 ($\pm$ 0.009)
1,024	0.95 ($\pm$ 0.013)	0.96 ($\pm$ 0.010)

E More Ablation Study

Sensitivity to initialization. To evaluate the sensitivity of our encoder to initialization, we repeated the hierarchical Riemannian Gaussian mixture fitting 100 times on 1,000 objects with randomly re-initialized root latents $\mathbf{z}$. As shown in Tab. 10, the NLL and IoU variance across different initializations is consistently small, indicating that our hierarchical fitting procedure is robust to the choice of starting point. We attribute this stability to the level-wise hierarchical likelihood objective, which provides smooth and well-conditioned gradients across all encoder layers.

Number of encoder levels and spliting manner. We ablate the impact of the number of hierarchical encoder levels on generation quality and intrinsic preservation. As shown in Tab. 11, a single-level encoder fails to capture part-level structure, yielding poor FID, KID, and c-mKNN. Performance improves consistently up to 3 levels, which balances semantic granularity with optimization stability. Adding a 4th level slightly degrades all metrics, likely due to over-segmentation of parts that leads to unstable gradient flows across levels and weakened semantic coherence at the coarsest level. We further ablate the impact of the splitting method of our encoder. As shown in Tab. 12, performing larger initial splits yields better fitting quality, while using fewer splitting layers further improves the overall fit.

Dimension of z. We ablate the effect of the root latent dimension on fitting quality. As shown in Tab. 10, lower dimensions severely limit the representational capacity of the hierarchical encoder, resulting in markedly worse NLL and IoU. Performance saturates beyond 768, while computational cost continues to grow. We therefore adopt $\mathbf{z} \in \mathbb{R}^{768}$ as the root latent throughout all experiments.

Asymmetric GW-inspired relational regularization. Tab. 13 extends the main-paper ablation by additionally reporting c-mKNN to measure intrinsic-preservation. Removing the regularization entirely collapses c-mKNN to 0.064, near the level of unstructured baselines, confirming that GW-inspired regularization is the primary driver of intrinsic preservation. The symmetric setting ($\alpha = \beta$) recovers partial structure but treats source and target neighborhoods equally, which over-constrains the geometric freedom on the 3D side and limits generation quality. The target-prioritized setting ($\alpha < \beta$) achieves higher c-mKNN than the symmetric case but degrades FID and KID compared to our source-prioritized design, because forcing the generated 3D to rigidly match its own neighborhood structure limits the global distributional quality. Our source-prioritized design ($\alpha > \beta$) achieves the best trade-off: it strongly penalizes violations of the semantic neighborhood in the condition space while allowing geometric flexibility in the 3D target space, yielding the best performance across all three metrics.

Table 11: Ablation on the number of encoder levels and the resulting generation quality and intrinsic preservation.

Levels	FID↓	KID↓	c-mKNN↑
1	17.83	4.51	0.073
2	14.57	3.26	0.101
3	**12.26**	**2.31**	**0.138**
4	13.04	2.69	0.126

Table 12: Ablation study on different encoder splits.

Split	NLL↓	IoU↑
[4,4,1]	3.93	0.71
[4,4,2]	3.06	0.79
[8,4,4]	**0.97**	**0.96**
[4,8,4]	1.15	0.92
[4,4,8]	1.24	0.91
[2,4,4,4]	1.35	0.89
[2,2,2,4,4]	1.52	0.86

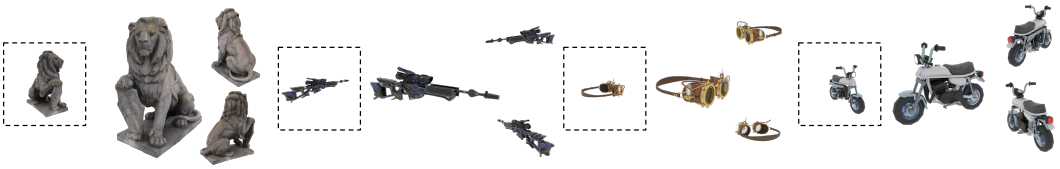

Figure 8: Additional visualization of Image-to-3D generation.

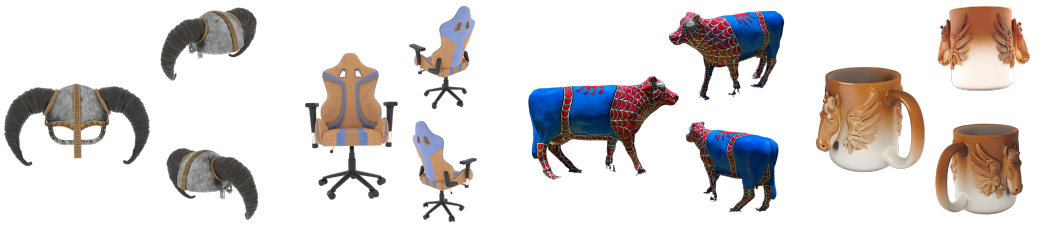

Figure 9: Additional visualization of Text-to-3D generation.

Necessity of SB formulation. Tab. 14 shows that SB formulation is essential because it provides a direct data-to-data transport mechanism with theoretical optimality guarantees. Diffusion models lack deterministic per-node endpoints, requiring full regeneration for each edit and leading to poorer preservation of unchanged regions. We further compare diffusion with GW regularization and CFG against our IPSB, where IPSB achieves substantially better generation quality, confirming the necessity of SB. Without SB, we cannot optimally transport part graphs to their corresponding 3D objects. Moreover, the Doob h-transform of SB provides a principled operator for our GW-inspired regularization, enabling direct structural constraints on part graphs during transport. Beyond generation, SB is also critical for editing. Intervention-based inference relies on pinned endpoints, and freezing non-target node trajectories in Alg. 1 is well-defined because each node-wise SB is conditioned on a fixed endpoint (Sec. 4.3).

Table 13: Extended ablation on asymmetric GW-inspired relational regularization, additionally reporting c-mKNN intrinsic-preservation.

Configuration	**FID↓**	**KID↓**	**c-mKNN↑**
No regularization	17.41	4.73	0.064
Symmetric ($\alpha = \beta$)	15.67	3.89	0.091
Target-prioritized ($\alpha < \beta$)	14.58	3.45	0.103
Source-prioritized ($\alpha > \beta$)	**12.26**	**2.31**	**0.138**

Table 14: Comparison of different optimization paradigms. Lower values indicate better performance.

Method	FID ↓	KID ↓
CFG + Diffusion	16.42	8.37
GW + Diffusion	10.81	4.95
GW + SB	**5.67**	**2.13**



Figure 10: Visualization of failure cases

F More Visualization and Experiments Analysis

F.1 Statistic analysis of evaluation.

Variance of main metrics. Tab. 15 reports mean and standard deviation of all primary generation metrics over 10 independent runs with different random seeds. The low variance of our method across all tasks confirms that IPSB training is stable.

Significance of c-mKNN improvements. The calibrated mKNN (c-mKNN) metric [16] measures how much the local neighborhood structure of a condition is preserved in the generated 3D output, after removing the bias introduced by trivial nearest-neighbor overlap via a permutation-based null model. In practice, c-mKNN values for cross-modal 3D generation are considerably lower than the theoretical maximum of 1.0, because the modality gap between 2D/text inputs and 3D outputs inherently limits perfect isometric transport. Empirically, a c-mKNN of $\approx$0.2 represents excellent intrinsic preservation for this task: it implies that, even after correcting for overlap bias, approximately 20% of local neighborhoods are consistently transferred from the condition to the generated 3D, a stringent requirement given the heterogeneity of the source and target spaces. Under this scale, the gap between our method and the strongest baseline represents a relative improvement, which is substantial and statistically significant. Moreover, the ablated variant without regularization demonstrates that virtually all of this gain is attributable to our asymmetric GW-inspired relational regularization, confirming that IPSB is the key component driving intrinsic-preserving transport.

F.2 Additional part-level evaluation and visualization.

To complement the object-level CLIP-S and human evaluations reported in main paper, we introduce a part-level CLIP Score (**part CLIP-S**) that measures semantic alignment at the granularity of individual parts rather than the whole object. Specifically, for each generated object we extract the rendering of each part segment (identified via the part graph), compute its CLIP embedding, and measure cosine similarity against the corresponding part-level conditioning input (*e.g.*, the CLIP embedding of the image or text prompt that guided that part). As shown in Tab. 16, our method achieves the highest part CLIP-S among all baselines, with a notable margin over both the ablated variant without GW regularization and the strongest baseline CoPart. This confirms that IPSB not only improves global generation quality but also achieves finer-grained semantic alignment at the

Table 15: Variance analysis of main generation metrics. Each result is the mean over 10 independent runs; numbers in parentheses indicate standard deviation.

Method	Image-to-3D			Text-to-3D		3D-to-3D	
	LPIPS↓	FID↓	KID↓	CLIP-S↑	Time(s)↓	FID↓	KID↓
Ours	**0.1021**(±.003)	**4.99**(±.21)	**1.98**(±.10)	**33.02**(±.16)	**2.12**(±.05)	**9.89**(±.29)	**6.24**(±.31)

Table 16: Part-level CLIP Score (part CLIP-S) evaluation. Scores are computed between the CLIP embedding of each generated part and its corresponding part-level conditioning input.

Metric	G.Cube	Photo3D	CoPart	TraCe	DiffGS	Ours w/o GW	**Ours**
Part CLIP-S↑	21.34	23.67	26.89	24.55	22.98	27.41	**30.76**

part level. We additionally provide new visualizations in Fig. 11 that illustrate the correspondence between each part-level condition and its generated counterpart, making the part-to-part matching induced by our encoder and IPSB directly interpretable.

F.3 Semantic Meaningfulness of Learned Parts

A part corresponds to a mixture component at the finest hierarchy level. Its latent can be decoded into mixture parameters or explicit 3D primitives. The encoder is trained with hierarchical likelihood, reconstruction, and rendering objectives, without 3D part annotations or segmentation labels. To complement the part CLIP-S evaluation, Tab. 17 reports segmentation mIoU on ShapeNetPart and PartNet. Our method outperforms the flat GMM and SATR baselines on both datasets, providing additional evidence that the learned components capture semantic part structure. The supervised PointNet++ results are included as a reference.

F.4 Graph Construction and Regularization

Graph construction. We compare cosine kNN, Euclidean kNN, and threshold-based graph construction in Tab. 18. Cosine kNN achieves the lowest FID and KID among the evaluated settings, supporting its use for constructing part graphs from semantic latents.

Alternative regularizers. Pointwise anchoring constrains individual latent coordinates, while Laplacian smoothing encourages neighboring latents to become similar. Neither directly specifies the edge-wise reference distances preserved by our GW-inspired regularization. By matching pairwise geodesic relations, our regularizer constrains the intermediate relational geometry without requiring a

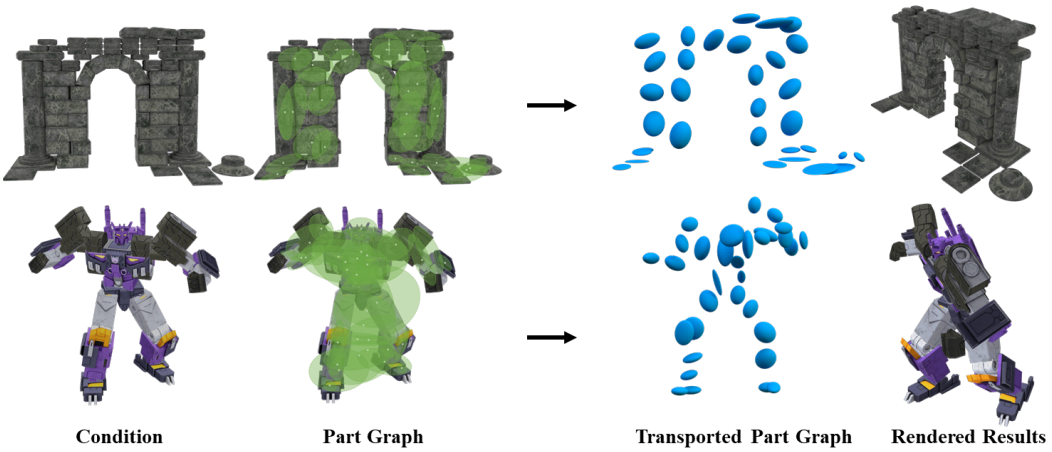

Figure 11: Visualization of part graph about condition and transported results.

Table 17: Part segmentation mIoU (%). PointNet++ uses supervised training and is shown separately.

Method	ShapeNetPart	PartNet
Flat GMM (k-means baseline)	21.4	19.7
SATR	31.9	24.2
Ours	**39.1**	**29.2**
PointNet++ (supervised)	81.9	48.6

Table 18: Comparison of graph construction strategies.

Graph construction	FID-50K ↓	KID-50K ↓
Cosine kNN	**5.16**	**2.08**
Euclidean kNN	13.84	6.91
Threshold graphs	18.27	9.53

Table 19: Comparison of regularization strategies.

Regularization	FID ↓	KID ↓
Pointwise anchor	18.73	9.46
Laplacian regularization	12.58	6.21
GW-inspired (ours)	**5.33**	**2.03**

Table 20: Image-to-3D FID under different input occlusion ratios. Lower is better.

Method	5%	10%	15%	20%	25%
CoPart	6.82	8.37	21.54	23.76	26.93
Ours	**5.16**	**5.41**	**5.78**	**6.24**	**6.91**

shared coordinate system across modalities. Tab. 19 shows that this design yields lower FID and KID than the two alternatives.

F.5 Robustness to Occlusion

The image encoder represents visible content and does not explicitly allocate a distinct component to a fully occluded part. Missing geometry is inferred from learned 3D priors and asymmetric relational regularization: target components without visible counterparts are generated under global conditioning, while the target-side term encourages structural consistency. Tab. 20 reports image-to-3D FID under increasing input occlusion. IPSB exhibits a smaller degradation than CoPart across the evaluated ratios.

F.6 Additional visualization of Image-to-3D and Text-to-3D.

We provide additional qualitative results for image-to-3D and text-to-3D generation in Fig. 8 and Fig. 9, respectively, further demonstrating the consistency and detail fidelity of our method across a diverse range of object categories and conditioning inputs.

F.7 Failure Cases

We present failure cases in Fig. 10, which primarily arise under extreme topology-changing articulations. Such conditions reduce the semantic discriminability of the part graphs constructed from the inputs, thereby degrading generation quality.

G Future Work

Extending IPSB to other 3D representations is a promising direction because transport operates on part graphs. For meshes, sampled vertices could be encoded with a spatial metric, followed by point

decoding and surface reconstruction. For SDFs, coordinate-latent pairs could provide the inputs, with a decoder trained through SDF or occupancy regression. For sparse voxels, occupied cells or octree leaves could be encoded using spatial and feature information, with occupancy predictions as decoder outputs. These extensions would require representation-specific input encodings, metrics, decoders, and reconstruction objectives, while retaining the part-graph transport formulation.

H Broader Impact

The proposed Intrinsic-Preserving Schrödinger Bridge (IPSB) framework advances direct part-aware 3D generation by enabling semantically consistent and topology-aware synthesis across multimodal conditions. By incorporating hierarchical part-level cross-modal encoder and asymmetric GW-inspired relational regularization, our method achieves high-fidelity 3D generation with precise part-level alignment and compositional controllability. This capability makes IPSB a versatile tool for applications in 3D content creation, virtual reality, robotics simulation, and digital design, where fine-grained control and structural consistency are essential for improving both usability and reliability.

However, similar to other generative 3D modeling approaches, the deployment of IPSB in real-world scenarios requires careful consideration, particularly in sensitive domains involving human-related assets or critical environments. Although our framework primarily operates on publicly available or synthetic datasets, reducing concerns related to personal data exposure, risks remain when models are applied to real-world inputs. Generated or edited 3D assets could potentially be misused for creating misleading virtual content, unauthorized replication of proprietary designs, or simulation environments that do not accurately reflect safety constraints. To mitigate such risks, it is important to enforce responsible usage practices, including controlled access to generation systems, clear guidelines for acceptable applications, and mechanisms for monitoring and auditing outputs. We emphasize the importance of aligning future research and deployment with ethical standards to ensure that advances in controllable 3D generation contribute positively to society.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The proofs appear in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the details in the main paper and supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code will be released upon acceptance; all datasets used are either publicly available or accessible upon request.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the details in the main paper and supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide the statistical significance in supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details in the supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that the paper adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide the Broader impacts analysis in the supplement.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We will release after acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper describes the human evaluation procedure and provides general instructions. Participants were compensated following standard practices.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The paper involves human annotators for evaluation purposes. We followed standard ethical guidelines for human evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The LLM is used only for editing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.