DreamBridge: BRIDGED CROSS-MODAL DISTILLATION FOR TEXT-TO-3D GENERATION

Anonymous authors
Paper under double-blind review

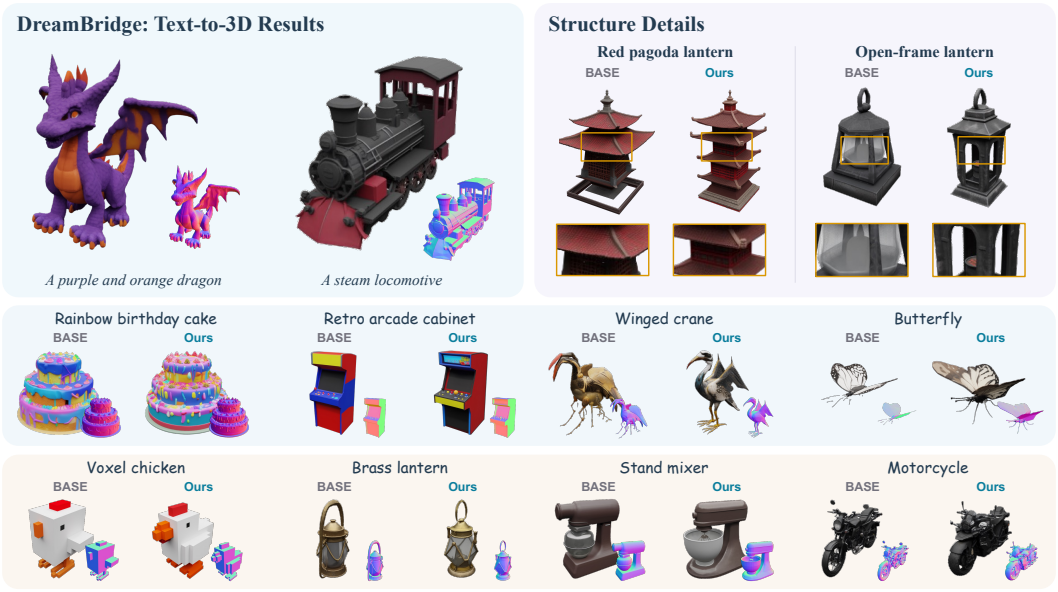

Figure 1: High-quality 3D assets generated by our method ***DreamBridge***, comparing with TRELLIS (baseline). The textual prompts are generated by GPT-6, as adopted by the baseline TRELLIS model.

ABSTRACT

Existing text-to-3D generation models can produce semantically consistent 3D objects from natural-language prompts, yet they still struggle to capture the geometry and spatial layout of complex objects. Common approaches either train native 3D generators on large paired datasets or combine text-to-image synthesis with image-to-3D generation at inference time. Inspired by the latter, we propose ***DreamBridge***, a bridged cross-modal distillation framework that transfers sparse-structure priors from a frozen image-to-3D teacher to a text-to-3D student, where synthesized images bridge the conditioning modalities during training. ***DreamBridge*** comprises three components: (i) *Bridged Triplets Synthesis*, which constructs triplets of a prompt, a filtered FLUX-generated image, and a cached teacher sparse-structure latent; (ii) *Native Latent Distillation*, which adapts only the student's sparse-structure flow with LoRA through conditional flow matching on cached teacher latents; and (iii) *Anchor Replay*, which supplements teacher supervision with structural samples from the frozen original text model. The overall adaptation procedure uses model-generated supervision without additional ground-truth 3D training data and retains text-only inference. On BridgeStruct-1.5K, our evaluation suite of 1,536 held-out prompts, ***DreamBridge*** improves structural agreement with the teacher over the pretrained baseline and achieves the highest reported text–3D semantic alignment scores among the evaluated native text-to-3D models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Generating 3D assets from natural-language descriptions is a central goal of generative 3D modeling. Early text-to-3D systems optimize a per-instance 3D representation against a frozen text-to-image diffusion model through score distillation (Poole et al., 2023; Wang et al., 2023), but per-prompt optimization makes them slow and prone to geometric artifacts. Recent native 3D generators avoid per-instance optimization by learning compact 3D latent distributions directly on large 3D corpora (Jun & Nichol, 2023; Lan et al., 2024; Chen et al., 2025; Xiang et al., 2024).

Image-to-3D generators offer a complementary route to 3D asset generation. Models such as TRELLIS-image (Xiang et al., 2024), TripoSG (Li et al., 2025b), and Step1X-3D (Li et al., 2025a) use visible shape and appearance cues in an input image to guide 3D generation with learned 3D priors. These explicit visual conditions can specify details that remain underspecified in text alone.

Two common strategies for improving text-to-3D generation are large-scale native-model training and cascaded generation. The first trains or finetunes text-to-3D models on paired text–3D data (Jun & Nichol, 2023; Lan et al., 2024), requiring high-quality 3D assets and text annotations. The second combines text-to-image synthesis with image-to-3D generation or multi-view reconstruction, as in pipelines based on LGM (Shi et al., 2024; Tang et al., 2024), InstantMesh (Xu et al., 2024), Direct2.5 (Lu et al., 2024), or Hunyuan3D-2 (Team, 2025b). These cascades provide explicit visual conditions for 3D generation but require additional image-generation stages at inference time.

To address these limitations, we propose ***DreamBridge***, a bridged cross-modal distillation framework that transfers the sparse-structure prior of a frozen image-to-3D teacher to a text-to-3D student. The image modality serves as a training-time *bridge*; the student uses only text at inference. ***DreamBridge*** instantiates this idea with three components (Figure 2). *Bridged Triplets Synthesis* (**BTS**) constructs each triplet from a prompt, a quality-gated synthesized image, and a sparse-structure latent sampled by the frozen teacher from that image. *Native Latent Distillation* (**NLD**) adapts the student with low-rank adapters confined to its sparse-structure flow, trained by conditional flow matching toward the teacher-generated targets in the shared native latent space. *Anchor Replay* (**AnR**) regularizes adaptation by incorporating latent samples from the frozen original text model's structural distribution. Our proposed ***DreamBridge*** uses only model-generated supervision, requires no additional ground-truth text–3D training pairs, and updates only the low-rank adapters in the sparse-structure flow.

We evaluate ***DreamBridge*** on BridgeStruct-1.5K, our text-to-3D evaluation suite of 1,536 held-out prompts, on 1,250 assets from the public Toys4K dataset, and on a fixed 1,000-asset subset of Google Scanned Objects (GSO) (Downs et al., 2022) under our evaluation protocols. On BridgeStruct-1.5K, ***DreamBridge*** improves structural agreement with the image-to-3D teacher over the pretrained baseline and achieves the highest reported text–3D semantic-alignment scores among the evaluated native text-to-3D models. Substantial occupancy gains are also observed in the 2K configuration, while the 8K configuration provides further occupancy improvements with a larger training budget. The main contributions of this paper are summarized as follows:

- We propose ***DreamBridge***, a bridged cross-modal distillation framework that transfers sparse-structure knowledge from an image-to-3D teacher to a text-to-3D student through a shared native latent space, while retaining text-only inference.
- Our ***DreamBridge*** is composed of *Bridged Triplets Synthesis*, *Native Latent Distillation*, and *Anchor Replay* for adapting the base models without additional ground-truth 3D training data. LoRA parameters for adaptation account for only 1.33% of the original sparse-structure flow model.
- On BridgeStruct-1.5K, ***DreamBridge*** improves teacher-referenced structural agreement over the pretrained baseline and achieves the highest reported text–3D semantic-alignment scores among the evaluated native text-to-3D models. Evaluation on Toys4K also shows modest improvements in text–3D semantic-alignment point estimates over the pretrained baseline.

2 RELATED WORK

Knowledge transfer for 3D generation. Knowledge distillation provides a general framework for transferring predictive or representative knowledge from a teacher model to a student

model (Hinton et al., 2014). Beyond settings where teacher and student share the same input modality, cross-modal knowledge transfer exploits complementary information across modalities to supervise models with different inputs. For 3D generation, this perspective is particularly relevant because image and text conditions provide different forms of information about object appearance and spatial structure from diverse modalities.

A prominent line of work transfers pretrained visual priors into 3D representations. Score Distillation Sampling and related methods (Poole et al., 2023; Wang et al., 2023; Tang et al., 2023) use pretrained diffusion models to guide per-instance optimization of 3D representations through differentiable rendering. Complementary approaches exploit multi-view diffusion, image inpainting, or video diffusion to provide image-based guidance for constructing 3D objects and scenes (Shi et al., 2024; Shriram et al., 2025; Yu et al., 2025). In these approaches, supervision is mediated by rendered images or synthesized views, typically involving per-instance optimization or multiple generation and reconstruction stages. Distillation has also been used to amortize 3D generation. ET3D (Chen et al., 2023) trains a text-to-3D generator using multi-view images from a pretrained diffusion teacher, while GECO (Wang et al., 2024) accelerates image-to-3D generation through two-stage distillation. In this paper, we transfer sparse-structure samples from a native image-to-3D teacher to a pretrained text-to-3D generator through their shared latent interface, with conditional flow matching and text-only inference.

Native 3D generation. Large-scale 3D corpora (Deitke et al., 2023b;a) have enabled generative models that learn distributions over native 3D representations using diffusion or flow objectives (Ren et al., 2024; Vahdat et al., 2022; Cheng et al., 2023). Recent systems extend this direction toward scalable models spanning diverse object categories and conditioning modalities (Chen et al., 2025; Team, 2025a; Xiang et al., 2024). Image-to-3D generation has also advanced substantially through both feed-forward reconstruction models and native 3D generators (Tang et al., 2024; Xu et al., 2024; Team, 2025a; Li et al., 2025b;a). Images provide explicit cues about visible shape, part arrangement, and appearance, although occluded geometry remains ambiguous.

Native text-to-3D models, including Shap-E (Jun & Nichol, 2023), LN3Diff (Lan et al., 2024), 3DTopia-XL (Chen et al., 2025), and the text branch of TRELLIS (Xiang et al., 2024), generate 3D representations directly from natural-language descriptions. Text conveys semantic concepts and relationships but often leaves their precise geometric realization underspecified. This ambiguity makes it challenging to recover fine structures and complex part arrangements from text alone. A common strategy is to first synthesize conditioning images and then apply image-to-3D generation or reconstruction (Tang et al., 2024; Xu et al., 2024). Such cascades exploit explicit visual conditions, but introduce additional models at inference time and can propagate errors from image synthesis into the resulting 3D assets. TIGON (Cen et al., 2026) studies joint text–image conditioning through cross-modal fusion during 3D generation. Our focus is complementary: ***DreamBridge*** uses synthesized images during training to transfer sparse-structure knowledge from an image-to-3D teacher to a text-to-3D student, retaining text-only inference.

3 PRELIMINARIES

Structured 3D generation. In this paper, we build ***DreamBridge*** on TRELLIS (Xiang et al., 2024), whose generation pipeline comprises sparse-structure generation, structured latent (SLat) generation, and decoding. A sparse-structure latent $z_0 \in \mathbb{R}^{8\times16\times16\times16}$ is decoded into active voxel coordinates, which provide the spatial support for SLat features. These features encode geometry and appearance and are mapped to a 3D asset by pretrained decoders. By default, we use TRELLIS-image-large as the teacher and initialize the student from TRELLIS-text-xlarge.

Interface for structure transfer. The two models use compatible sparse-structure representations and a shared sparse-structure decoding interface. Teacher-generated latents can therefore serve directly as targets for the student's structure flow. We choose this interface because it determines the spatial support on which subsequent 3D generation operates, allowing adaptation to focus on structural layout while reusing the pretrained SLat generator and decoders. For a prompt p, let $c(p)$ denote its frozen text encoding and $v_{\theta_0,\phi}$ the student's structure velocity field, with frozen pretrained parameters θ_0 and trainable LoRA parameters ϕ. Throughout, $t \in [0, 1]$ denotes normalized flow time; the backbone's fixed time-input rescaling is absorbed into the notation for $v_{\theta_0,\phi}$.

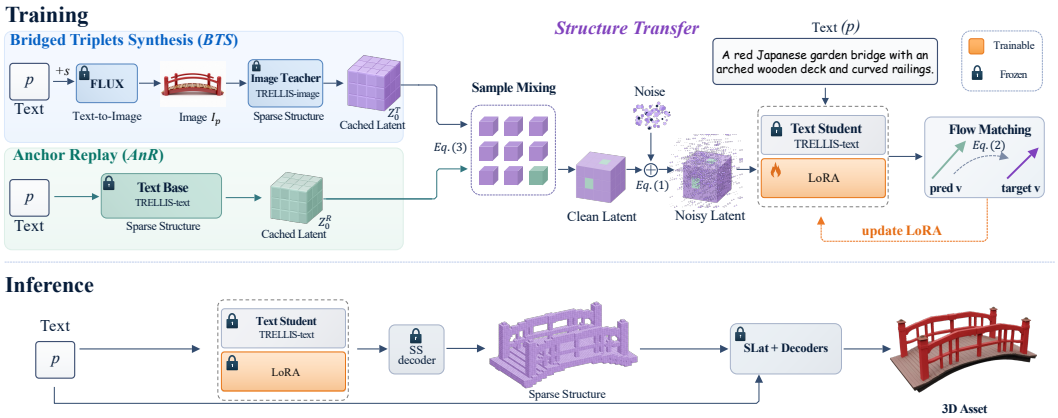

Figure 2: **Overview of *DreamBridge*.** (**Top**) Synthesized images bridge text prompts to a frozen image-to-3D teacher, which supplies cached structure targets. LoRA adapts the student's sparse-structure flow using teacher targets and frozen-base replay. (**Bottom**) At inference, the adapted flow and frozen downstream stages generate 3D assets directly from text.

4 METHODOLOGY

Our method ***DreamBridge*** combines offline supervision construction with adaptation in the shared sparse-structure latent space (Fig. 2). *Bridged Triplets Synthesis* constructs text–image–structure triplets using an image-to-3D teacher (Sec. 4.1). *Native Latent Distillation* uses the resulting text–latent pairs to train the student's structure flow through low-rank adapters with standard conditional flow matching (Sec. 4.2). *Anchor Replay* supplements the teacher targets with samples from the frozen original text model, incorporating its pretrained structural distribution into the adaptation objective (Sec. 4.3). The adapted structure generator then operates as a pure text-to-3D pipeline during inference time.

4.1 BRIDGED TRIPLETS SYNTHESIS (BTS)

Image bridge. We construct prompts from fixed vocabularies and templates spanning object categories, text lengths, and structural complexity levels. For each prompt p, FLUX.1-dev (Labs, 2024) generates candidate images from $p \oplus s$, where $\oplus$ denotes concatenation and s is a fixed presentation suffix specifying object framing and background. We select an accepted image I_p using the fixed ranking and recovery procedure in Appendix C.2; prompts without an accepted image do not enter the teacher-target pool. The student is conditioned on the original prompt p, and the suffix is used only for image synthesis.

Native structural targets. Let $\mathcal{A}_G$ denote the image-generation and acceptance procedure above, and let $\mathcal{T}_{\mathrm{SS}}$ denote the frozen image-to-3D teacher's native sparse-structure sampler. For each image-accepted prompt p, we construct

$$I_p = \mathcal{A}_G(p \oplus s), \qquad z^T_{0,p,k} = \mathcal{T}_{\mathrm{SS}}(I_p; \xi_{p,k}),$$

where $\xi_{p,k}$ is the teacher sampling noise and k indexes the teacher samples for p. The synthesized *bridge triplet* $(p, I_p, z^T_{0,p,k})$ connects the original text, generated image, and teacher-generated structure. Each $z^T_{0,p,k}$ is a terminal sparse-structure sample, cached before sparse-structure decoding and paired with the original prompt p; the student uses only $(p, z^T_{0,p,k})$ during optimization. Fixed latent and decoded-occupancy checks determine target validity. We write z^T_0 for a retained teacher sample when the prompt and sample indices are omitted. Each eligible teacher prompt contributes a set of valid cached samples, which are cycled through during training. These cached text-latent pairs from our proposed *bridged triplets synthesis (BTS)* supervise the student in Sec. 4.2. All targets are generated offline, so the image generator and teacher are unnecessary during student optimization.

4.2 NATIVE LATENT DISTILLATION (NLD)

Parameter-efficient structural adaptation. We insert LoRA adapters (Hu et al., 2021) into the self-attention, cross-attention, and feed-forward linear layers of the student's sparse-structure Transformer. For each adapted weight, $W' = W + (\alpha/r)BA$, where W is frozen, A and B are trainable low-rank factors, r is the adapter rank, and α controls the update scaling. Optimization updates only the adapter parameters ϕ. The text encoder, sparse-structure decoder, SLat generator, and 3D decoders remain frozen during the training time.

Learning from cached structure samples. We train the adapted velocity field with the standard conditional flow-matching objective used by the pretrained backbone (Lipman et al., 2023; Xiang et al., 2024). The bridge changes the source of structural supervision: cached samples from the image-to-3D teacher are paired with the original text prompts, while the flow-matching formulation is retained. For a cached terminal sample z_0, draw normalized time $t \sim p_{\text{time}}$, where p_{time} is a logit-normal sampling distribution, and noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Following the pretrained model's flow direction, the interpolation path and its target velocity are

$$\begin{aligned} x_t &= (1-t)z_0 + \big[\sigma_{\min} + (1-\sigma_{\min})t\big]\epsilon, \\ u_t &= \frac{\mathrm{d}x_t}{\mathrm{d}t} = (1-\sigma_{\min})\epsilon - z_0, \end{aligned} \tag{1}$$

where $\sigma_{\min}$ is a small positive noise floor. Thus, $t = 1$ is the noise endpoint and $t = 0$ corresponds to $z_0 + \sigma_{\min}\epsilon$.

To support classifier-free guidance (Ho & Salimans, 2021), we sample $m \sim \text{Bernoulli}(\delta_c)$, where δ_c is the text-condition dropout probability. We use the empty-prompt encoding when $m = 1$ and the original text encoding otherwise, yielding the condition $\widetilde{c}(p; m)$. The loss for a cached pair is

$$\ell_\phi(z_0, p) = \mathbb{E}_{t,\epsilon,m}\left[\frac{1}{D}\left\|v_{\theta_0,\phi}\big(x_t, t, \widetilde{c}(p;m)\big) - u_t\right\|_2^2\right], \tag{2}$$

where D is the number of elements in the sparse-structure latent. The teacher supplies terminal structure samples; the velocity target u_t is computed analytically from each cached sample and newly drawn noise using Eq. (1), without querying the teacher's velocity field during training. Thus, distillation here is sample-based: teacher-generated terminal samples provide supervision for the text-to-3D student.

4.3 ANCHOR REPLAY (ANR)

Replay from the text prior. Inspired by generative replay in continual learning (Wu et al., 2018), we retain samples from the pretrained text prior while learning from new teacher supervision. Teacher targets reflect the image-conditioned generation process and can differ from samples of the original text model. For the same pool of image-accepted prompts, we independently sample terminal latents z_0^R from the frozen original text model conditioned on p. Replay targets follow the latent and occupancy checks and cache-cycling procedure described in Sec. 4.1. Validity is assessed separately for the two sources, so their eligible prompt sets may differ. These samples supplement the teacher-derived supervision with training targets from the student's pretrained text prior, providing a source of regularization during adaptation.

Mixed training objective. Let $\mathcal{D}_T$ and $\mathcal{D}_R$ denote the sampling distributions over valid cached teacher and replay pairs, respectively. The schedule balances object categories and text-length bins while using a fixed mixture of structural-complexity levels, with greater weight assigned to complex prompts. We apply the same sample loss from Eq. (2) to both sources and minimize the following objective with respect to ϕ:

$$\begin{aligned} \mathcal{L}(\phi) = (1-\rho)\,\mathbb{E}_{(p,z_0^T)\sim\mathcal{D}_T}\big[\ell_\phi(z_0^T, p)\big] \\ + \rho\,\mathbb{E}_{(p,z_0^R)\sim\mathcal{D}_R}\big[\ell_\phi(z_0^R, p)\big], \end{aligned} \tag{3}$$

where ρ denotes the replay fraction. Source frequencies implement the mixture weights, with a uniform mean loss over each global batch. Exact allocation and source-specific validity handling are detailed in Appendix B.3.

4.4 TEXT-TO-3D INFERENCE

The adapted student retains the original text-to-3D generation pipeline. Given a prompt p, it samples a sparse-structure latent from Gaussian noise using the original text-conditioned sampler and classifier-free guidance, integrating from $t = 1$ toward $t = 0$. The frozen sparse-structure decoder, SLat generator, and 3D decoders then produce the asset as described in Sec. 3, with SLat generation conditioned on the same prompt. The resulting text-to-3D pipeline operates only from text, without the participation of the image as the intermediate bridge modality.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUPS

Evaluation Protocol. We primarily evaluate on *BridgeStruct-1.5K*, a frozen 1,536-prompt test suite held out from training, covering six splits: in-distribution, novel nouns, novel templates, short-and-simple, complex-structure, and prior-regression prompts. All methods use the same prompts and a unified 12-view rendering protocol, with metrics aggregated at the prompt level. Model selection is performed solely on a disjoint validation set. We further evaluate on the public *Toys4K* (Stojanov et al., 2021) benchmark and a fixed 1,000-asset subset of *Google Scanned Objects (GSO)* (Downs et al., 2022) under dataset-specific protocols. Detailed validity handling, paired-comparison rules, and external-dataset settings are provided in Appendix D.

Evaluation Metrics. On BridgeStruct-1.5K, occupancy IoU/F1, Chamfer distance, normal consistency, depth error, and silhouette IoU evaluate structural agreement with Teacher-IT; the main table reports cyclically aligned silhouette IoU. On Toys4K and GSO, geometric metrics instead use ground-truth 3D assets. For text–3D alignment, CLIP (Radford et al., 2021) measures multi-view image–text similarity ($\times 100$), while Uni3D (Zhou et al., 2024) and ULIP-2 (Xue et al., 2024) report point–text cosine similarity. CLIP is averaged over 12 views on BridgeStruct-1.5K and GSO and eight views on Toys4K. Full metric definitions are provided in Appendix D.

Comparison Methods. We compare ***DreamBridge*** with seven representative text-to-3D systems in two groups. *End-to-end native models* directly generate 3D representations from text, including Shap-E (Jun & Nichol, 2023), the official text-to-3D version of LN3Diff (Lan et al., 2024), and 3DTopia-XL (Chen et al., 2025). *Cascade pipelines* first generate images or multi-view observations from text and then reconstruct a 3D representation. These include LGM with MVDream (Tang et al., 2024; Shi et al., 2024), InstantMesh with FLUX (Xu et al., 2024; Labs, 2024), Direct2.5 (Lu et al., 2024), and Hunyuan3D-2 (Team, 2025b). Method-specific settings and representation proxies are detailed in Appendix D.1.

All systems use matched prompts, rendering protocols, and metric implementations where applicable. BASE is the original text-to-3D TRELLIS model (Xiang et al., 2024). The two frozen teacher references share image-conditioned sparse structure: *Teacher-IT* uses text-conditioned SLat, as do BASE and ***DreamBridge***, whereas *Teacher-II* uses image-conditioned SLat and forms the full image-to-3D pipeline. The suffixes denote SS/SLat conditioning, respectively. Teacher-IT provides the structural reference; neither teacher is a universal upper bound on the metrics.

Implementation Details. Bridged Triplets Synthesis (BTS) uses entirely synthetic training data. For each text prompt, the frozen FLUX.1-dev model (Labs, 2024) generates an image, and the frozen image-to-3D TRELLIS teacher converts the accepted image into a sparse-structure target. This process yields 8,062 filtered prompt–image pairs and requires no additional real image–caption pairs or ground-truth 3D assets. In Native Latent Distillation (NLD), we start from the text-to-3D TRELLIS model (Xiang et al., 2024) and optimize only rank-16 LoRA adapters (Hu et al., 2021) in the sparse-structure flow through conditional flow matching on cached teacher latents. The adapters contain 13.0M trainable parameters, or 1.33% of the original flow model; all remaining parameters are frozen. We train for 12,000 steps using AdamW (Loshchilov & Hutter, 2017) with a global batch size of 32 and a learning rate of 10^{-4}. Anchor Replay (AnR) mixes teacher targets with samples from the frozen text-to-3D base model on a fixed 9:1 source schedule. At inference, the adapted model retains the original text-to-3D generation pipeline and frozen appearance stage. Additional data-generation and optimization details are provided in the supplementary material.



Figure 3: **Qualitative comparisons under matched text prompts.** Color renderings are paired with translucent normal renderings to reveal part arrangements and open regions. LGM+MVDream has color renderings only.

5.2 Experimental Results

With the protocol fixed, we first inspect generated structures and then compare the complete model combining BTS, NLD, and AnR quantitatively. Prompt-level and external-dataset analyses follow the aggregate comparison.

Qualitative Comparison. Figure 3 compares ***DreamBridge*** with the evaluated baselines under matched text prompts. The examples, including a rocking horse and an open crown, illustrate differences in geometric detail, part arrangement, and open regions. In the shown rocking-horse example, ***DreamBridge*** preserves visibly separated curved runners and connecting components. The crown example also shows distinct arches and intervening gaps, illustrating the open structures targeted by the adaptation.

Quantitative Comparison. Table 1 summarizes results on BridgeStruct-1.5K, Toys4K, and GSO. On BridgeStruct-1.5K, ***DreamBridge*** achieves higher point estimates than BASE for silhouette agreement and all three text–3D semantic-alignment metrics. It achieves the highest reported scores among the evaluated end-to-end text-to-3D systems on this benchmark. Compared with cascade pipelines, it obtains higher silhouette IoU and ULIP-2 scores, while InstantMesh and Hunyuan3D-2 retain advantages in CLIP and Uni3D, respectively. The native-branch metrics in Table 2 also show higher occupancy IoU, F1, and normal consistency, and lower Chamfer distance than BASE, although depth error does not improve. These gains are obtained with text-only inference, without an upstream image generator, multi-view model, or per-instance optimization.

Prompt-level and Split-wise Analysis. Figure 4 reports prompt-level paired differences with pointwise 95% bootstrap intervals: 1,518 valid pairs for SS F1 and 1,519 for CLIP-12. Coverage, aggregation, and resampling details are given in Appendix D.7.

Figure 4 examines how the aggregate improvement is distributed across prompts and evaluation splits. In the right-hand scatter plot, 81.8% of the valid SS F1 pairs lie above the diagonal, indicating higher teacher-referenced SS F1 under ***DreamBridge***. Panel (a) shows positive mean SS F1 differences

Table 1: **Quantitative comparisons on BridgeStruct, Toys4K, and GSO.** The best and second-best results within each category are highlighted in bold and underlined, respectively. Greener cells indicate better values within each column. Teacher-II is excluded from ranking.

		BridgeStruct-1.5K (1,536)				Toys4K (1,250)					GSO (1,000)				
Method	E2E	Sil. IoU↑	CLIP↑	Uni3D↑	ULIP-2↑	CD-24↓	NC-24↑	CLIP↑	Uni3D↑	ULIP-2↑	CD-24↓	NC-24↑	CLIP↑	Uni3D↑	ULIP-2↑
LGM	×	0.5093	30.01	0.1136	0.1859	**0.0193**	0.5346	28.30	0.1893	0.2576	0.0227	0.5332	31.00	0.1098	0.1519
InstantMesh	×	**0.5569**	**30.29**	0.1752	0.1881	0.0199	**0.6507**	**29.58**	0.2031	0.2769	0.0252	0.6326	**32.48**	0.1623	0.1919
Direct2.5	×	0.4648	27.35	0.1410	0.1531	0.0439	0.6357	27.49	0.2141	**0.2908**	0.0340	**0.6434**	28.68	0.1375	0.1777
Hunyuan3D-2	×	0.4093	26.37	**0.2117**	**0.2326**	0.0327	0.5320	21.62	**0.2335**	0.2551	**0.0208**	0.6342	31.68	**0.1714**	**0.2057**
Shap-E	✓	0.4019	27.38	0.1654	0.1787	0.0133	0.6820	30.29	0.2065	0.2629	0.0200	**0.6593**	30.44	0.1543	0.1827
LN3Diff	✓	0.4056	24.01	0.1129	0.1249	0.0312	0.6088	22.41	0.1279	0.1526	0.0232	0.6087	27.17	0.1093	0.1262
3DTopia-XL	✓	0.3467	25.68	0.1484	0.1471	0.0192	0.6476	26.10	0.2131	0.2599	0.0207	0.6536	28.20	0.1477	0.1549
BASE (TRELLIS)	✓	0.5012	29.79	0.1899	0.2299	**0.0118**	**0.6900**	32.14	0.2321	0.2850	0.0196	0.6398	31.32	0.1596	0.1814
DreamBridge	✓	**0.5717**	**30.21**	**0.1930**	**0.2388**	0.0119	0.6898	**32.17**	**0.2331**	**0.2888**	**0.0189**	0.6493	**31.34**	**0.1624**	**0.1917**
Teacher-II	–	0.9507	30.67	0.1918	0.2500	0.0144	0.6682	32.27	0.2210	0.2878	0.0184	0.6516	32.24	0.1656	0.2048



Figure 4: **Prompt-level and split-wise comparisons on BridgeStruct-1.5K. (Left)** Mean differences between ***DreamBridge*** and BASE in teacher-referenced SS F1 (a) and CLIP-12 (b) across six test splits. Error bars indicate pointwise 95% paired-bootstrap confidence intervals, and n denotes the number of valid paired prompts. ***DreamBridge*** achieves higher mean SS F1 across all six splits. **(Right)** Each point compares the teacher-referenced SS F1 scores of ***DreamBridge*** and BASE for the same prompt. The dashed diagonal indicates equal scores; 81.8% of valid paired prompts lie above it, showing improved structural agreement with the teacher for a majority of evaluated prompts.

across all six splits, with each pointwise interval above zero and the largest mean gain on the complex-structure split. Structural improvements therefore extend across multiple prompt categories and a majority of evaluated prompts. Panel (b) shows positive mean CLIP-12 differences across the splits, although the interval for prior regression includes zero. The semantic improvement on this split remains inconclusive, despite its positive structural gain.

Evaluation on Toys4K. Moving from teacher-referenced BridgeStruct comparisons to GT geometry, the Toys4K panel of Table 1 evaluates ***DreamBridge*** on 1,250 assets from the public Toys4K dataset (Stojanov et al., 2021) under our evaluation protocol. Text–3D semantic-alignment point estimates show modest improvements over BASE, while the GT-aligned CD-24 and NC-24 scores remain close to those of BASE and slightly favor the baseline. Because the text prompts, rendering setup, and metric implementations are specific to our evaluation protocol, the absolute scores should not be directly compared with results obtained under different protocols.

Evaluation on GSO. As a second GT-referenced check, Table 1 reports a title-conditioned evaluation on 1,000 fixed GSO assets. On the 999 valid BASE/***DreamBridge*** pairs, ***DreamBridge*** has higher NC-24 and ULIP-2 point estimates. Selection, paired uncertainty, and evaluation details are provided in Appendix D.5.

Table 2: Comparison of training-data scales.

	SS IoU↑	SS F1↑	CD↓	Normal↑	Depth↓	CLIP-12↑
BASE	0.0848	0.1512	0.0783	0.5610	0.1719	29.79
DreamBridge-2K	0.1110	0.1930	0.0635	0.5999	**0.1577**	29.98
DreamBridge-8K	**0.1203**	**0.2069**	**0.0600**	**0.6151**	0.1744	**30.21**
Teacher-II	1.0000	1.0000	0.0000	0.9484	0.0145	30.67

5.3 ABLATION STUDIES

We ablate the Anchor Replay, the replay fraction, and the training-data scale as below. The ablation results of Presentation-suffix are provided in Appendix E.3.

Anchor Replay and Replay Fraction. An earlier 2K development experiment compares 10% replay (R10) with teacher-only adaptation (R0) at matched training updates. At 5,000 steps, R10 improves teacher-referenced SS F1 from 0.1580 to 0.1602 and the base-relative text CFG-direction cosine from 0.5431 to 0.5785 (Table 3); the same directions are observed at 4,000 steps. These results support using replay to retain the pretrained text-conditioned response while learning teacher structures. In a separate replay-fraction comparison, R10 yields the largest SS IoU improvement and Chamfer reduction among R10, R25, and R40. We therefore use 10% replay under the structure-prioritized validation rule. The full development results are provided in Appendix E.2.

Table 3: Anchor Replay (5K development).

Variant	SS F1↑	CFG cosine↑
Teacher-only (R0)	0.1580	0.5431
10% replay (R10)	**0.1602**	**0.5785**

Training-data Scale. We trained ***DreamBridge*** on 2K and 8K prompt-image pairs, while keeping the NLD objective, AnR schedule, and LoRA setup unchanged. As shown in Table 2, ***DreamBridge***-2K already improves SS F1 from 0.1512 to 0.1930, while ***DreamBridge***-8K further increases it to 0.2069. These results show that substantial structural gains are already obtained with the smaller training set, with the 8K configuration achieving higher structural agreement with the teacher.

6 CONCLUSION

We presented ***DreamBridge***, a bridged cross-modal distillation framework for transferring structural knowledge from an image-to-3D teacher to a text-to-3D student. *Bridged Triplets Synthesis* links text prompts, synthesized images, and teacher-generated sparse-structure latents. *Native Latent Distillation* adapts the student's structure flow with LoRA using these native 3D targets, while *Anchor Replay* supplements them with samples from the frozen original text model. The framework requires no additional ground-truth 3D training data and retains fully text-only inference. Experiments on BridgeStruct-1.5K demonstrate improved structural alignment with teacher-generated targets over the pretrained baseline and leading text–3D semantic alignment among the evaluated native text-to-3D models. These findings support model-generated structural targets as an effective source of supervision for adapting pretrained 3D generators across conditioning modalities. On the title-conditioned GSO evaluation, normal consistency and ULIP-2 improve over the baseline on valid pairs. More broadly, they suggest that the structural knowledge of image-to-3D models can benefit text-to-3D generation without requiring image synthesis at inference time. The current evidence does not establish improved ground-truth geometry on Toys4K or transfer beyond TRELLIS. Extending this transfer from sparse structure to appearance generation is a promising direction for future work.

Limitations. ***DreamBridge*** is currently developed and validated primarily on the TRELLIS backbone, and adapting it to other native text-to-3D architectures may require backbone-specific modifications. Its training relies on synthetic supervision produced by an image generator and an image-conditioned 3D teacher, and therefore remains subject to the quality and coverage of these pretrained components.

AI USE STATEMENT

We used generative AI tools to improve the language and readability of parts of this manuscript and to generate the text prompts shown in Figure 1. Generative AI tools also assisted in constructing the prompts in the BridgeStruct-1.5K evaluation dataset. In our training pipeline, FLUX.1-dev generates synthetic images, while pretrained image-to-3D and text-to-3D models provide structural targets and replay samples. Generative text-to-3D and image-to-3D systems are also evaluated as baselines, as described in the paper. The authors reviewed the final manuscript and take responsibility for its text, claims, experimental results, and AI-assisted artifacts.

REFERENCES

Jiazhong Cen, Jiemin Fang, Sikuang Li, Guanjun Wu, Chen Yang, Taoran Yi, Zanwei Zhou, Zhikuan Bao, Lingxi Xie, Wei Shen, and Qi Tian. Text-image conditioned 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yiming Chen, Zhiqi Li, and Peidong Liu. Et3d: Efficient text-to-3d generation via multi-view distillation, 2023. URL https://arxiv.org/abs/2311.15561.

Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26576–26586, 2025.

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4456–4465, 2023.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023a.

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*, pp. 13142–13153, 2023b.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022. URL https://arxiv.org/abs/2204.11918.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NIPS Deep Learning Workshop*, 2014.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *ECCV*, 2024.

Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025a.

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*, 2025b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2.5: Diverse text-to-3d generation via multi-view 2.5d diffusion. *Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

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*, pp. 8748–8763. PMLR, 2021.

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4209–4219, 2024.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *International Conference on Learning Representations*, 2024.

Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *3DV*, 2025.

Stefan Stojanov, Anh Thai, and James M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias, 2021. URL https://arxiv.org/abs/2101.07296.

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.

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. *arXiv preprint arXiv:2402.05054*, 2024.

Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025a.

Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025b.

Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in neural information processing systems*, 35:10021–10039, 2022.

Chen Wang, Jiatao Gu, Xiaoxiao Long, Yuan Liu, and Lingjie Liu. Geco: Generative image-to-3d within a second, 2024. URL https://arxiv.org/abs/2405.20327.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36:8406–8441, 2023.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Chenshen Wu, Luis Herranz, Xialei Liu, Joost Van De Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems*, 31, 2018.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27091–27101, 2024.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024.

A STATEMENT

Broader Impacts. Our proposed ***DreamBridge*** contributes to improving the quality and controllability of text-to-3D generation, potentially lowering the technical barrier to creating 3D content from natural-language descriptions. Such capabilities can support a broad range of applications, including game development, film and virtual production, digital design, education, cultural heritage visualization, robotics simulation, and immersive content creation, benefiting both professional creators and users with limited 3D modeling expertise. At the same time, increasingly capable text-to-3D systems may facilitate the generation of assets that imitate copyrighted designs, branded products, or other protected visual content, raising potential intellectual property and misuse concerns. Moreover, generated assets may reflect ambiguities or biases present in the input descriptions and underlying training data. Responsible deployment, appropriate content safeguards, and clear usage policies are therefore important for reducing potential risks while enabling beneficial applications.

Ethical Statement. We acknowledge the potential ethical implications associated with text-to-3D generative models, including risks related to intellectual property, inappropriate content generation, and misuse of synthetic 3D assets. Because text prompts provide flexible and open-ended control over generated content, the resulting assets may reproduce undesirable biases or generate content that resembles protected or sensitive visual concepts. The datasets, pretrained models, and other external resources used in this work are employed in accordance with their respective licenses and usage policies. To promote transparency and responsible research, we will provide the implementation and experimental details necessary to reproduce our method and evaluation. We encourage researchers and practitioners to consider the provenance of training data, generated content, and downstream application contexts when developing or deploying text-to-3D systems.

Reproducibility Statement. To facilitate reproducibility, we will release the source code, training and inference scripts, model configurations, and evaluation scripts required to reproduce the experimental results reported in this paper after the peer review process. Where permitted by the licenses of the underlying models and datasets, we will also provide the trained model checkpoints and relevant preprocessing utilities. All experiments are conducted using publicly accessible datasets and established text-to-3D backbones. Detailed descriptions of the model architecture, training configuration, inference settings, data preparation, sampling procedures, and evaluation protocols are provided in the main paper and appendices.

B IMPLEMENTATION DETAILS

This appendix specifies the offline supervision pipeline, the trainable components, and the evaluation procedures used by ***DreamBridge***. The adaptation data are generated by pretrained models: an image generator provides visual conditions, an image-conditioned 3D teacher provides native sparse-structure targets, and the original text-conditioned model provides replay targets. No additional ground-truth 3D assets or real image–caption pairs are used for this adaptation. This statement concerns our adaptation stage, rather than the data used to pretrain the underlying models.

B.1 MODEL COMPONENTS AND ADAPTER PLACEMENT

TRELLIS-text-xlarge initializes the student; TRELLIS-image-large supplies the structural teacher. FLUX.1-dev supplies the synthetic images used to condition the teacher. The student and teacher expose a compatible sparse-structure (SS) interface, with terminal latents of shape $8 \times 16 \times 16 \times 16$. Targets are cached immediately after SS sampling and before the SS decoder. The frozen decoder converts these latents into active voxel coordinates. At inference, these coordinates condition the frozen structured-latent (SLat) stage and its decoders, using the original text condition.

Only LoRA parameters in the student's SS flow are trainable. Each adapted linear layer has the form

$$W' = W + \frac{\alpha}{r} BA, \qquad r = 16, \quad \alpha = 16, \tag{4}$$

where W remains frozen. LoRA dropout is zero. The SS transformer has 28 blocks, and each block contains the seven adapted projections listed in Table 4. The resulting 196 adapted linear modules introduce 12,959,744 trainable parameters. Relative to the 975,440,648 parameters of the original SS

Table 4: LoRA placement in each SS transformer block. Module names identify the implemented projections.

Component	Adapted projections
Self-attention	`self_attn.to_qkv`, `self_attn.to_out`
Cross-attention	`cross_attn.to_q`, `cross_attn.to_kv` `cross_attn.to_out`
Feed-forward network	`mlp.mlp.0`, `mlp.mlp.2`

flow, this is approximately 1.33%. The denominator is the SS flow, rather than the entire 3D pipeline. The text encoder, original SS weights, SS decoder, SLat flow, and final decoders are frozen.

B.2 Flow Time, Conditional Targets, and Loss Normalization

We distinguish normalized flow time from the numerical time input expected by the pretrained network. If $f_{\theta_0,\phi}$ denotes that network, the velocity field used in the main text is

$$v_{\theta_0,\phi}(x,t,c) = f_{\theta_0,\phi}(x, 1000t, c), \qquad t \in [0,1]. \tag{5}$$

This conversion affects the time embedding only. The interpolation and target velocity are defined with respect to t, so the target velocity is not multiplied by 1000.

For either a teacher or replay terminal latent z_0, the training draws are

$$a \sim \mathcal{N}(1,1), \qquad t = \mathrm{sigmoid}(a), \qquad \epsilon \sim \mathcal{N}(0,I), \qquad \sigma_{\min} = 10^{-5}, \tag{6}$$

where the normal distribution for a has mean one and standard deviation one. The conditional path and its analytical velocity are

$$x_t = (1-t)z_0 + [\sigma_{\min} + (1-\sigma_{\min})t]\epsilon, \tag{7}$$

$$u_t = (1-\sigma_{\min})\epsilon - z_0. \tag{8}$$

Thus $x_1 = \epsilon$ and $x_0 = z_0 + \sigma_{\min}\epsilon$; inference integrates from the noise endpoint toward smaller t. The small endpoint noise is part of the implemented path, even though z_0 itself is a cached terminal sample.

Independently of the target source, the text condition is replaced with the frozen empty-prompt encoding with probability 0.1. Let $\widetilde{c}(p)$ denote the resulting condition. The sample loss is

$$\ell(z_0,p) = \frac{1}{D}\left\| f_{\theta_0,\phi}(x_t, 1000t, \widetilde{c}(p)) - u_t \right\|_2^2, \qquad D = 8 \cdot 16^3. \tag{9}$$

The network observes x_t, t, and the text condition; it is not given z_0 as a separate input. Here "conditional" flow matching refers to the endpoint-conditioned training path and should not be confused with text conditioning. Teacher endpoints are sampled offline. Training does not query the teacher velocity at the student's intermediate states.

B.3 Exact Replay Allocation and Optimization

The replay fraction is $\rho = 0.1$. With a global batch of 32, one allocation cycle spans ten updates and contains 320 sample positions: 288 teacher targets and 32 replay targets. Source assignments are shuffled within each cycle, so a batch may contain both sources. The protocol does not consist of nine pure teacher batches followed by one pure replay batch. Each update averages Eq. 9 uniformly over its 32 positions. The mixture weights are implemented by source frequencies; no additional source-specific loss multiplier is applied.

The prompt schedule balances eight categories and three length bins. Simple, moderate, and complex prompts receive proportions of 25%, 25%, and 50%, respectively. Teacher and replay availability are checked independently. Invalid targets are excluded from their own source pool, and a missing teacher target is not replaced with a replay target, or vice versa. Cached targets are cycled according to the fixed schedule.

Table 5: Optimization settings for the 8K and 2K training configurations.

Setting	Full adaptation	Smaller-data comparison
Accepted prompt–image pairs	8,062	2,048
Global batch	32	32
Optimizer	AdamW	AdamW
Peak learning rate	10^{-4}	10^{-4}
Weight decay	0	0
Warmup updates	400	100
Post-warmup schedule	Cosine	Cosine
Gradient norm clipping	1.0	1.0
Forward autocast	bfloat16	bfloat16
Adapter parameter precision	float32	float32
Checkpoint interval	1,000 updates	500 updates

The distributed implementation uses a microbatch of one per device and gradient accumulation to obtain the stated global batch. The mean squared error is computed in float32. The smaller-data experiment uses comparable teacher-target exposure per eligible prompt. Because its training duration also changes, the comparison does not isolate dataset size at fixed compute.

B.4 Sampling and Decoding

Both sparse-structure (SS) and structured-latent (SLat) stages use `FlowEulerGuidanceIntervalSampler` with $\sigma_{\min} = 10^{-5}$. For N steps, define $u_i = 1 - i/N$ and $t_i = 3u_i/(1 + 2u_i)$, $i = 0, \ldots, N$. At each step the velocity network receives the time input $1000t_i$ and the Euler update is

$$x_{i+1} = x_i + (t_{i+1} - t_i)\,\widehat{v}(x_i, t_i). \tag{10}$$

Within the configured guidance interval, $\widehat{v} = (1 + w)v_{\text{cond}} - wv_{\text{empty}}$; outside it, $\widehat{v} = v_{\text{cond}}$.

For target caching, the image-conditioned SS teacher uses 25 steps, $w = 5.0$, $t \in [0.5, 1.0]$, and time-rescaling factor 3.0. The frozen text SS replay sampler uses 25 steps, $w = 7.5$, $t \in [0.5, 0.95]$, and the same time rescaling. For native BridgeStruct-1.5K text-only inference, both SS and SLat use the latter 25-step text configuration. The Toys4K BASE/***DreamBridge*** evaluation uses the separately frozen override of 50 steps and $w = 3.0$ for both SS and SLat; its remaining sampler settings follow the text pipeline. Neither image synthesis nor the image-conditioned teacher is invoked at student inference time.

The frozen SS decoder maps logits greater than zero to active coordinates on a 64^3 grid. The text-conditioned SLat sampler operates on those coordinates; its output is unnormalized using the frozen model's SLat mean and standard deviation before mesh and Gaussian decoding. Meshes are exported to PLY with `trimesh.Trimesh(process=False)`; when vertex attributes contain RGB values, their first three channels are clipped to $[0, 1]$ and converted to 8-bit vertex colors. Gaussians are saved as PLY with no export transform. No additional mesh simplification is applied by this export path.

C Synthetic Supervision and Benchmark Construction

C.1 Prompt Construction

Training prompts are constructed from fixed vocabularies and templates covering eight groups: organic objects and characters; products and instruments; furniture and architecture; vehicles and machinery; thin rods, spokes, gears, cages, and chains; holes, hollow structures, grids, and negative space; asymmetric, articulated, and multipart objects; and transparent, reflective, and material-oriented descriptions.

Prompt length is measured by the CLIP tokenizer without special tokens. The short, medium, and long bins contain 8–24, 25–42, and 43–60 tokens. The three complexity levels use one, two, or four

Table 6: Accepted BridgeStruct training prompts (C01–C08). Each sample ID is an exact row identifier in the frozen 8,062-sample training manifest.

Category	Tier	Sample ID	Original student prompt
C01	simple	`flux_train_c01_0038`	Carved walnut crane sculpture; low oval body, thin wings.
C01	moderate	`flux_train_c01_0270`	Smoked glass crane sculpture; broad winged form, four legs, thin wings.
C01	complex	`flux_train_c01_0525`	Smoked glass rabbit puppet; tapered body, offset tail, hinged neck, side crest, shell plates.
C02	simple	`flux_train_c02_0002`	Clear glass rotary mixer; vertical body, offset knobs.
C02	moderate	`flux_train_c02_0282`	Dark walnut manual typewriter; long body, offset knobs, four feet.
C02	complex	`flux_train_c02_0602`	Polished brass coffee mill; long body, inner chamber, open slot, two handles, curved spout.
C03	simple	`flux_train_c03_0022`	Smoked glass bridge table; stepped mass, thin rail.
C03	moderate	`flux_train_c03_0278`	White marble canopy chair; open framework, four legs, frame layers.
C03	complex	`flux_train_c03_0518`	Cast concrete wall niche; stepped mass, lower opening, side arches, frame layers, offset panel.
C04	simple	`flux_train_c04_0050`	Copper alloy crane truck; tracked base, central cabin.
C04	moderate	`flux_train_c04_0322`	Laminated wood tracked excavator; upright chassis, central cabin, hinged arm.
C04	complex	`flux_train_c04_0570`	Dark titanium survey rover; long chassis, four wheels, central cabin, side pontoons, rear engine.
C05	simple	`flux_train_c05_0042`	Carbon fiber wire cage; triangular frame, three rings.
C05	moderate	`flux_train_c05_0298`	Clear acrylic spiral spring; offset cage, three rings, four rods.
C05	complex	`flux_train_c05_0542`	Silver aluminum rod truss; narrow framework, four rods, two chains, side braces, three rings.
C06	simple	`flux_train_c06_0010`	Cast bronze open lantern; hollow body, side opening.
C06	moderate	`flux_train_c06_0274`	Translucent acrylic portal frame; hollow body, inner cavity, three grids.
C06	complex	`flux_train_c06_0578`	Dark stone open lantern; domed shell, inner cavity, thin rim, three grids, open underside.
C07	simple	`flux_train_c07_0042`	Painted steel modular creature; folded silhouette, central joint.
C07	moderate	`flux_train_c07_0318`	Aged brass offset turbine; folded silhouette, raised front, central joint.
C07	complex	`flux_train_c07_0613`	Aged brass modular creature; offset stance, three limbs, raised front, offset weight, rear pod.
C08	simple	`flux_train_c08_0002`	Glazed porcelain amber figurine; rounded silhouette, four lobes.
C08	moderate	`flux_train_c08_0306`	Mirror chrome acrylic chair; flowing form, material shells, thin rim.
C08	complex	`flux_train_c08_0514`	Iridescent enamel translucent helmet; translucent shell, hollow core, two clear layers, open loop, thin rim.

object features, respectively, together with one global shape constraint. These feature counts are construction rules, rather than measured levels of geometric difficulty. They do not establish that every longer or more complex prompt is harder for every method. Each category has a frozen lexicon of 16 in-distribution head nouns, eight held-out head nouns, eight materials, six global shapes, and eight local features. Four template families are used within each length bin: training and validation use T families, the held-out-template split uses separate U families, and a confirmation-only split uses G families. For medium prompts, the generator appends one fixed finish/tone sentence; for long prompts, it appends three fixed finish/tone/texture sentences. The complete frozen lexicons and template strings are listed below, and Table 6 gives one actual accepted training prompt for every category–complexity combination. The sample indices in those tables identify rows of the frozen 8,062-sample training manifest.

The construction audit requires uniqueness of full prompts and semantic structure keys across generated splits. The held-out-noun split has no head-noun or synonym-family overlap with training and validation, and the held-out-template split has no template-family overlap with them. For cross-split near-duplicate screening, candidate pairs sharing at least 0.55 of the smaller trigram set are compared by a normalized `SequenceMatcher` ratio and rejected at a ratio of 0.92 or higher. The frozen prompt and split audits record no violations. These checks concern text identity and near duplication; they do not prove that the held-out prompts are uniformly more difficult.

Frozen category lexicons.

C01 *In-distribution nouns:* heron automaton; fox marionette; octopus figurine; beetle sculpture; dragon statuette; dancer figure; tortoise model; moth ornament; centaur idol; mantis automaton; seahorse figure; crane sculpture; lizard totem; owl reliquary; rabbit puppet; crab figurine.

Held-out nouns: pangolin maquette; nautilus effigy; axolotl icon; ibex miniature; cuttlefish carving; cassowary doll; lemur statue; trilobite replica.

Materials: verdigris brass; painted ceramic; carved walnut; smoked glass; blackened steel; ivory resin; copper enamel; frosted acrylic.

Global shapes: arched torso; tapered body; low oval body; S curved profile; broad winged form; coiled upright body.

Local features: four legs; thin wings; open ribs; offset tail; three fins; hinged neck; side crest; shell plates.

C02 *In-distribution nouns:* gooseneck kettle; survey compass; concert violin; portable lantern; bench microscope; sewing machine; folding camera; desk fan; coffee mill; mantel clock; film projector; hand drill; table gramophone; manual typewriter; brass telescope; rotary mixer.

Held-out nouns: thumb kalimba; marine sextant; pyramid metronome; mandoline slicer; medical otoscope; rolling planimeter; counting abacus; wax phonograph.

Materials: brushed steel; polished brass; cream enamel; dark walnut; cast aluminum; clear glass; red bakelite; hammered copper.

Global shapes: vertical body; long body; stepped cylinder; boxy housing; tapered cone; triangular frame.

Local features: two handles; curved spout; three dials; open slot; four feet; hinged arm; offset knobs; inner chamber.

C03 *In-distribution nouns:* canopy chair; pedestal table; writing desk; arched bookcase; pavilion model; spiral cabinet; staircase console; balcony cabinet; curved chaise; room divider; tripod stool; wall niche; tower shelf; vaulted bench; screened cupboard; bridge table.

Held-out nouns: clerestory kiosk; campanile cabinet; cantilever settee; rotunda shelf; trefoil credenza; loggia screen; belfry console; apse cupboard.

Materials: oiled oak; black walnut; powdered steel; white marble; woven rattan; cast concrete; aged bronze; smoked glass.

Global shapes: tall rectangle; low cantilever; arched frame; stepped mass; open framework; narrow tower.

Local features: four legs; side arches; upper shelf; three bays; offset panel; thin rail; lower opening; frame layers.

C04 *In-distribution nouns:* survey rover; cargo tricycle; twin rotorcraft; harbor tugboat; steam locomotive; tracked excavator; swept glider; touring motorcycle; crane truck; research submarine; wind turbine; counterweight catapult; propeller sled; cable car; armored crawler; orbital lander.

Held-out nouns: single monowheel; flapping ornithopter; rail draisine; foil hydrocraft; tracked snowcat; open autogyro; portal gantry; deep bathyscaphe.

Materials: riveted steel; painted aluminum; dark titanium; weathered brass; laminated wood; matte composite; copper alloy; white ceramic.

Global shapes: long chassis; upright chassis; tapered fuselage; tracked base; cylindrical body; triangular frame.

Local features: four wheels; two rotors; open cockpit; three axles; hinged arm; rear engine; side pontoons; central cabin.

C05 *In-distribution nouns:* wire cage; spoked wheel; chain loop; gear train; antenna array; whisk frame; lattice globe; pulley rig; rod truss; bicycle wheel; link bracelet; gyroscopic cage; sprocket cluster; spiral spring; spoke fan; cable reel.

Held-out nouns: tensegrity prism; armillary sphere; chainmail torus; ferris rotor; wireframe dodecahedron; reed skeleton; filament harp; pinion carousel.

Materials: thin brass; black steel wire; silver aluminum; copper rod; carbon fiber; white ceramic; blue titanium; clear acrylic.

Global shapes: circular framework; narrow framework; spherical frame; triangular frame; nested rings; offset cage.

Local features: twelve spokes; three rings; six gears; two chains; four rods; empty hub; hinged loop; side braces.

C06 *In-distribution nouns:* perforated shell; hollow torus; lattice vessel; keyhole panel; grid dome; arch ring; tunnel block; vented cylinder; honeycomb screen; open lantern; pierced drum; slotted sphere; ribbed basket; portal frame; annular disc; mesh urn.

Held-out nouns: mobius sieve; oculus cube; catenoid shell; pierced tetrahedron; toroidal grille; vault skeleton; fenestrated cone; reticulated pod.

Materials: matte ceramic; cast bronze; frosted glass; brushed aluminum; dark stone; ivory resin; copper mesh; translucent acrylic.

Global shapes: hollow body; open rectangle; domed shell; annular form; perforated body; nested shell.

Local features: large through hole; six apertures; open underside; inner cavity; three grids; thin rim; opposing portals; side opening.

C07 *In-distribution nouns:* hinged robot; offset tripod; multipart lamp; articulated arm; asymmetric drone; modular creature; pivoted sculpture; segmented tool; cantilever device; folded walker; branched mechanism; nested carrier; split body vehicle; rotating assembly; jointed puppet; offset turbine.

Held-out nouns: scissor jack; pantograph beast; folding exoskeleton; differential walker; gimbal creature; bifurcated automaton; knuckle crane; eccentric rover.

Materials: painted steel; aged brass; dark hardwood; white ceramic; clear acrylic; black titanium; copper enamel; woven composite.

Global shapes: asymmetric body; split body; branching body; offset stance; folded silhouette; multipart frame.

Local features: three limbs; two hinges; offset weight; four modules; pivoted arm; raised front; rear pod; central joint.

C08 *In-distribution nouns:* crystal decanter; glass jellyfish; chrome vessel; translucent helmet; acrylic chair; mirrored sculpture; porcelain lamp; resin shell; frosted clock; polished kettle; iridescent beetle; glazed vase; obsidian mask; enamel fan; amber figurine; silver orb.

Held-out nouns: dichroic prism; alabaster censer; mercury droplet; opaline shrine; nacre reliquary; quartz carafe; ice goblet; lacquer monolith.

Materials: clear crystal; dichroic glass; mirror chrome; translucent resin; glazed porcelain; polished obsidian; iridescent enamel; frosted acrylic.

Global shapes: rounded silhouette; faceted silhouette; translucent shell; reflective body; hollow profile; flowing form.

Local features: two clear layers; hollow core; three facets; thin rim; opaque insert; four lobes; open loop; material shells.

Shared modifiers and template grammar. The shared finish words are satin, brushed, polished, weathered, glazed, frosted, hammered, matte; the tone words are cobalt, amber, ivory, charcoal, verdigris, crimson, silver, smoke gray; and the texture phrases are fine grained, subtly speckled, smooth banded, lightly mottled, evenly striated, softly translucent, mirror smooth, gently patinated. The four in-distribution short template strings are:

1. `One {material} {noun}: {shape}{tail}.`
2. `A {material} {noun} with {shape}{tail}.`
3. `{material} {noun}; {shape}{tail}.`
4. `Single {material} {noun}: {shape}{tail}.`

The four in-distribution medium/long template strings are:

1. `Build one {noun} from {material} around {shape}{tail}.`
2. `A single {noun}, made of {material}, follows {shape}{tail}.`
3. `In {material}, one complete {noun} is organized as {shape}{tail}.`
4. `The object is one {material} {noun}; its structure follows {shape}{tail}.`

The four held-out short template strings are:

1. `From {shape}: one {noun} in {material}{tail}.`
2. `Around {shape} stands a {material} {noun}{tail}.`
3. `With {shape}, form one {noun} from {material}{tail}.`
4. `As {shape}: a single {material} {noun}{tail}.`

The four held-out medium/long template strings are:

1. `Around {shape}, {material} forms one {noun}{tail}.`
2. `Defined first by {shape}, the object is a {material} {noun}{tail}.`

3. `With {shape} as its outer organization stands one {material} {noun}{tail}.`
4. `Organized around {shape} is a single {noun} in {material}{tail}.`

The confirmation-only short template strings are:

1. `Held-out {material} {noun}: {shape}{tail}.`
2. `One unseen {noun} in {material}: {shape}{tail}.`
3. `Confirmation {material} {noun}: {shape}{tail}.`
4. `New {material} {noun} with {shape}{tail}.`

The confirmation-only medium/long template strings are:

1. `For confirmation only, one {material} {noun} has {shape}{tail}.`
2. `A held-out {material} {noun} is built around {shape}{tail}.`
3. `This confirmation {noun} in {material} uses {shape}{tail}.`
4. `One unseen {material} {noun} combines {shape}{tail}.`

The tail is the comma-separated list of local features, or empty if no features are present. The four optional length-padding sentences are "Its {finish} finish has restrained {tone} variation.", "Broad faces remain {texture}.", "Narrow parts keep the same {finish} finish.", and "The full silhouette retains the {tone} response."

C.2 Image Generation and Acceptance

FLUX.1-dev produces 1024×1024 images using 28 inference steps, guidance strength 3.5, maximum sequence length 512, and bfloat16 inference. Its input is the original prompt followed by a fixed presentation suffix. The suffix specifies a consistent object presentation and is used only for image generation; the student and replay branch receive the original prompt. It is a fixed formatting convention in this pipeline. The recorded suffix is:

> one complete object, centered, fully visible, isolated neutral background, no crop, no pedestal, no text, no watermark, soft studio lighting

The frozen generation string is the unmodified student prompt, one space, this literal suffix, and a final period. The suffix is absent from the student and replay text conditions.

For each prompt, two primary FLUX candidates are generated first. Two reserve candidates are generated only when both primary candidates fail the frozen 2D acceptance gates. Within this initial pool, all valid candidates are ranked by the frozen input-image quality key: original- prompt/image CLIP cosine, largest foreground-region fraction, retained foreground fraction, lower visible-text proxy margin, and finally lower candidate index. For prompts unresolved after the initial pool, each of two recovery rounds supplies four ordered slots; the first passing slot in a round is selected. Thus a prompt that remains unresolved has at most 12 candidate attempts (four initial plus four per recovery round). No teacher or student 3D output is used for image selection. The initial prompt pool contains 8,192 items. The final accepted set has 8,062 images (98.413%): 7,872 from the initial pool, 132 from the first recovery round, and 58 from the second; the remaining 130 prompts are excluded from adaptation and retained for coverage diagnostics. Image acceptance is separate from SS target validity. Table 7 lists the numerical gates.

Image-gate operators. The original RGB image is scored by the frozen OpenAI CLIP ViT-B/32 processor and encoder against the unmodified student prompt; image and text embeddings are ℓ_2-normalized before their cosine is taken. The same image is processed by U2-Net background removal. Alpha values greater than 0.8×255 define the foreground mask. Foreground components use eight-neighbor connectivity; regions with fewer than $\max(16, \lfloor 0.001HW \rfloor)$ pixels are excluded from the component count, where H and W are the image dimensions. The largest-region fraction is

Table 7: Frozen numerical checks for synthetic image acceptance; the corresponding operators are defined in the text.

Check	Accepted range or threshold
Foreground fraction	$[0.015, 0.92]$
Connected foreground regions	At most 4
Largest-region fraction	At least 0.82
Preprocessing retained fraction	At least 0.97
Absolute area-change fraction	At most 0.75
Original-prompt/image CLIP cosine	At least 0.20
Visible-text/watermark proxy margin	At most 0.08

the area of the largest counted region divided by the total foreground area. A foreground box within $\max(2, 0.005 \max(H, W))$ pixels of any image border fails the no-crop check.

The preprocessing crop is a square centered on the foreground box with side length 1.2 times its longer side. The retained fraction is the foreground area within the image-clipped crop divided by total foreground area. To estimate whether background removal changes object area excessively, a raw-image foreground proxy uses Euclidean RGB distance greater than 28 from the median border color, followed by 3×3 binary opening and 5×5 closing. The reported area change is $(f_{\text{removed}} - f_{\text{raw}})/f_{\text{raw}}$; its absolute value is gated. Decode failure, an empty foreground mask, or border contact also rejects the candidate.

The visible-text/watermark score is a CLIP proxy rather than OCR: it is the image cosine with "a product image with visible text, letters, logo, caption, or watermark" minus the image cosine with "a clean studio product image with no visible text, letters, logo, or watermark". After selection, the official TRELLIS image preprocessing repeats background removal, crops the object using the same 1.2 box factor, and resizes the crop to 518×518 before conditioning the teacher.

C.3 Native Target Caching and Validity

The image teacher processes each accepted image and samples an SS latent. Separately, the frozen text base model samples a replay latent from the original prompt. Latents and text conditions are cached in float32; text conditions have shape 77×768. Cache validation checks the latent shape and finiteness, nonempty and unique voxel coordinates, and an active voxel count between 256 and 260,000. Active coordinates are obtained by thresholding SS decoder logits at zero.

The final source-specific masks contain 8,061 eligible teacher prompts and 8,058 eligible replay prompts. The excluded target sets are disjoint. The remaining exclusions are associated with more than 32 connected components. Source eligibility is retained explicitly throughout sampling. Components are counted on the 64^3 binary occupancy grid using six-neighbor, face-adjacent connectivity (a three-dimensional SciPy connectivity-1 structuring element and `ndimage.label`). The component limit is 32. One teacher target and four replay targets exceed it; their sample sets do not overlap. Accordingly, the two source pools remain separately masked rather than being replaced by their intersection.

Each per-sample cache record binds the sample ID, source and variant, frozen seed, initial-noise SHA-256, sampler parameters, model revision, decoder hash, SS-latent shape and finite-value statistics, active-voxel and component counts, and SHA-256 plus byte length for saved latent, coordinate, and condition files. Image-teacher records additionally bind the selected FLUX image and preprocessed image hashes. The latent is stored as float32 with shape $[1, 8, 16, 16, 16]$; decoded coordinates have shape $[N, 4]$ in batch–x–y–z order. Nonfinite latents, empty or repeated coordinates, and active-voxel counts outside $[256, 260000]$ invalidate only the affected source record.

C.4 Validation and Test Membership

BridgeStruct-1.5K contains 1,536 test prompts in six fixed splits (Table 8). Checkpoint selection uses a disjoint active validation set of 128 prompts. Teacher-dependent validation metrics use 126 prompts

Table 8: Membership of the frozen BridgeStruct-1.5K test suite. The validation set is separate and is not included in these counts.

Split	Prompts
In-distribution	512
Novel nouns	256
Novel templates	256
Short and simple	128
Complex structure	256
Prior-regression	128
Total	1,536

with valid teacher references, while teacher-independent semantic validation retains all 128. Missing teacher-dependent values remain unavailable rather than being assigned a numerical score.

Candidate checkpoints are evaluated at the intervals in Table 5. The selection rule prioritizes category-macro occupancy F1, followed by lower native Chamfer distance, higher occupancy IoU, greater validation CLIP improvement over BASE, a smaller active-voxel tail, and an earlier checkpoint. Eligibility also requires complete checkpoint and validation records and the fixed validity and stability checks. The selected full-adaptation checkpoint is at update 12,000; the smaller-data endpoint is at update 3,000. Test metrics are not used for checkpoint selection.

We evaluate a frozen deterministic subset of 1,250 Toys4K assets Stojanov et al. (2021). The official metadata lists 3,229 assets; all pass path, nonzero-byte, SHA-256, and Blender-header checks. Of these, 3,180 have a nonempty first caption and are eligible. Each eligible asset is ordered by the SHA-256 of a fixed selection namespace, a null separator, and its asset SHA-256, with the asset hash as a tie breaker; the first 1,250 form the evaluation manifest. The 49 assets without a usable first caption are ineligible, while the other 1,930 eligible assets are outside this fixed subset rather than quality-filtered. The evaluation prompt is the official metadata's first caption with surrounding whitespace removed, `captions[0].strip()`, and BASE and ***DreamBridge*** share an asset-specific deterministic inference seed.

The selected assets span 105 top-level archive categories. The largest are chair (77 assets), robot (45), dog (43), and bottle (32); Table 9 gives the complete distribution. The source assets are official `.blend` files. The geometry scorer uses the corresponding official 10,000-point GT point clouds as its reference; the GT silhouette renderer uses the Blender assets under the frozen camera and normalization protocol described below. This is our protocol-aligned evaluation on a public dataset, not a claim to recover an unpublished split or evaluator.

D EVALUATION PROTOCOL

D.1 COMPARED SYSTEMS AND REFERENCE DEFINITIONS

BASE is the original TRELLIS-text model from which the student is initialized. Native text-to-3D comparisons include Shap-E, the official text-conditioned LN3Diff model, and 3DTopia-XL. The evaluated cascades are LGM with MVDream, InstantMesh with FLUX, Direct2.5, and Hunyuan3D-2. The model inputs and evaluation identities are matched across methods where a metric is applicable.

The two teacher references share image-conditioned sparse structure. *Teacher-IT* uses text conditioning for SLat, whereas *Teacher-II* uses image conditioning for SLat. The two letters refer, in order, to the conditioning modalities of the SS and SLat stages. On BridgeStruct-1.5K, Teacher-IT provides the structural reference and uses the same downstream conditioning modality as the student. Teacher-II is an unranked reference pipeline. Neither is a universal upper bound on every metric.

The Hunyuan3D-2 row has different configurations in the BridgeStruct and Toys4K panels. On BridgeStruct-1.5K, it uses the FR40K configuration with Paint. On Toys4K, it uses the STANDARD shape-only configuration, with fixed gray appearance and no Paint. The latter therefore does not evaluate the full textured Hunyuan pipeline. LGM also requires representation-specific treatment:

Table 9: Complete category distribution of the frozen 1,250-asset Toys4K subset. Categories are the top-level directories of the official asset identifiers.

Category	n	Category	n	Category	n	Category	n
airplane	15	apple	19	ball	12	banana	9
bicycle	5	boat	11	bottle	32	bowl	10
bread	13	bunny	10	bus	7	butterfly	10
cake	19	candy	16	car	12	cat	24
cells_battery	9	chair	77	chess_piece	16	chicken	12
closet	1	coin	14	cookie	9	cow	6
crab	10	cup	18	cupcake	8	deer_moose	20
dinosaur	26	dog	43	dolphin	8	donut	10
dragon	15	drum	9	elephant	11	fan	7
fish	17	flower	22	fox	25	fridge	7
fries	8	frog	18	giraffe	7	glass	10
grapes	5	guitar	21	hamburger	7	hammer	27
hat	18	helicopter	6	helmet	8	horse	13
ice_cream	15	key	12	keyboard	3	knife	8
ladder	4	laptop	5	light_bulb	3	lion	6
lizard	7	marker	2	microwave	2	monitor	2
monkey	7	motorcycle	7	mouse	10	mug	24
mushroom	9	octopus	14	orange	7	panda	6
pc_mouse	1	pear	8	pen	3	pencil	9
penguin	23	phone	5	piano	3	pig	11
pizza	9	plate	10	radio	14	robot	45
sandwich	5	saw	2	screwdriver	6	shark	10
sheep	14	shoe	14	sink	8	snake	10
sofa	13	spade	6	stove	7	submarine	4
toaster	10	tractor	7	train	10	trashcan	13
tree	17	truck	9	tv	3	violin	5
whale	11						

its Toys4K point metrics use Gaussian proxies, as detailed below. These qualifications apply to the corresponding rows even when compact main-table labels omit them.

D.2 Rendering and View Aggregation

For BridgeStruct-1.5K, each generated asset is rendered from 12 uniformly spaced azimuth views, $\{0°, 30°, \ldots, 330°\}$, at a fixed elevation of $20°$. All RGB views are rendered at a resolution of 512×512 on a white background, with a camera distance of 2.0 and a field of view of $40°$. For text–image semantic evaluation, the similarity scores are first averaged over the 12 rendered views for each prompt and then aggregated across prompts. If a maximum-over-views score is additionally reported, it is explicitly labeled and does not replace the mean-view metric.

For Toys4K, each asset is rendered from eight uniformly spaced azimuth views, $\{0°, 45°, \ldots, 315°\}$, at a fixed elevation of $30°$. The rendering protocol uses a resolution of 512×512, a camera distance of 2.0, a field of view of $40°$, and a white background. External meshes are centered using their axis-aligned bounding boxes and uniformly scaled such that the longest bounding-box dimension is one, without applying any GT-dependent orientation selection. Native outputs and GT meshes retain their predefined canonical orientations. Available vertex or texture colors are preserved during rendering without additional relighting. For representation-specific cases, LGM uses its native Gaussian RGB values, while the shape-only Hunyuan configuration is rendered with a fixed gray value of $100/255$.

D.3 Teacher-Referenced Sparse Structure and Silhouettes

Let A and B be the generated and teacher active-voxel sets in the shared grid. We compute

$$\mathrm{IoU}(A,B)=\frac{|A\cap B|}{|A\cup B|},\qquad \mathrm{F1}(A,B)=\frac{2|A\cap B|}{|A|+|B|}. \tag{11}$$

Empty or invalid outputs are handled by the validity protocol, rather than silently assigned a perfect score. These quantities measure agreement with the teacher structure, not ground-truth reconstruction accuracy.

The cross-system silhouette measure in the main table is derived from white-background RGB renders. A pixel is foreground if any RGB channel is below 255. The binary mask is BOX-downsampled from 512×512 to 128×128 and thresholded at 0.5. It is compared with the corrected teacher mesh reference mask. This RGB-derived measure is distinct from the native geometry-renderer silhouette metric.

For generated masks M_v and reference masks R_v, index alignment uses the same azimuth index. Cyclic alignment selects one common circular shift for the full view sequence:

$$S_{\mathrm{index}}=\frac{1}{12}\sum_{v=0}^{11}\mathrm{IoU}(M_v,R_v), \tag{12}$$

$$S_{\mathrm{cyclic}}=\max_{k\in\{0,\dots,11\}}\frac{1}{12}\sum_{v=0}^{11}\mathrm{IoU}(M_{(v+k)\bmod 12},R_v). \tag{13}$$

The main comparison reports the cyclic measure. It compensates for a discrete azimuth-phase difference; it is not a learned pose-alignment component of ***DreamBridge*** and does not resolve arbitrary orientation errors.

Native geometry metrics supplement SS overlap. The native coordinate Chamfer is the half-sum of bidirectional mean *unsquared* Euclidean nearest-neighbor distances between active coordinates in the 64^3 grid; when a unit-grid value is reported, this voxel distance is divided by 63. A separate mesh-surface Chamfer uses 10,000 area-weighted samples from each generated and teacher mesh and the same unsquared distance. Neither native Chamfer definition is the squared, normalized Toys4K CD.

For each of 12 matching native views, let M_v and R_v denote the generated and teacher foreground masks and $J_v=M_v\cap R_v$. Views with an empty mask, empty intersection, or empty union are invalid. Normal agreement averages the cosine between decoded unit normals over J_v where both normal magnitudes exceed 10^{-6}; depth error averages $|d_v^{\mathrm{gen}}-d_v^{\mathrm{teacher}}|$ over finite shared pixels and divides by the teacher foreground depth range for that view. The metric is valid only when all 12 views have usable normals, depths, and nondegenerate teacher depth ranges. Per-view values are then averaged uniformly; missing views are not silently discarded.

D.4 Toys4K Geometry and the 24-Rotation Alignment

The geometry evaluation uses normalized point clouds with 4,000 farthest-point-sampled points. For point sets P and Q, the reported Chamfer distance is

$$\mathrm{CD}(P,Q)=\frac{1}{2}\left[\frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2^2+\frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|q-p\|_2^2\right]. \tag{14}$$

No additional scale factor is applied. Normal consistency is the mean absolute cosine of corresponding nearest-neighbor normals, symmetrized over the two point sets.

Let $\mathcal{R}_{24}$ denote the 24 proper rotations of a cube: signed axis-permutation matrices with determinant $+1$. The aligned result uses

$$R^*=\arg\min_{R\in\mathcal{R}_{24}}\mathrm{CD}(RP,Q),\quad \text{CD-24}=\mathrm{CD}(R^*P,Q),\quad \text{NC-24}=\mathrm{NC}(R^*P,Q). \tag{15}$$

Generated normals are rotated together with their points. Crucially, NC-24 uses the CD-minimizing rotation; it does not independently select the rotation with the highest normal score. This alignment

uses GT geometry and is a post-hoc evaluation procedure. It is not available to the generator at inference, and it is not applied to text-semantic scores. Raw geometry results retain the original orientation.

LGM does not provide mesh-surface points under this protocol. Its geometry uses opacity–area-weighted Gaussian centers and covariance-derived normals. These are representation proxies, and their distances and normal scores should not be interpreted as measurements of an identically sampled triangle-mesh surface. LGM has no Toys4K mesh-silhouette value in this evaluation. For Toys4K mesh geometry, the reference is the official point-cloud archive. Generated meshes are sampled at 10,000 area-weighted surface points with corresponding face normals. Each point set is separately centered and scaled to unit maximum radius, with normals normalized to unit length; PointNet++ farthest-point sampling reduces each set to 4,000 points before CD and NC.

D.5 GSO Selection and Evaluation

The GSO panel is an author-defined evaluation on a fixed set of 1,000 assets selected from the 1,033 release IDs using a deterministic SHA-256 ordering. The input prompt is constructed from the first nonempty official description line and the category when available. This title-conditioned prompt need not identify an instance's exact shape. The aggregate results are reported as descriptive comparisons.

The GT reference is each selected GSO mesh. For BASE and ***DreamBridge***, one generated asset per prompt is evaluated with the same frozen checkpoints and sampling configuration: 50 SS and 50 SLat steps, with CFG 3.0 in each stage. Each GT and generated mesh is sampled at 10,000 area-weighted surface points and reduced to 4,000 points by farthest-point sampling after independent centroid and unit-radius normalization. CD-24 and NC-24 use the same squared-distance and CD-selected proper-rotation definitions as Eq. 14; unlike Toys4K, the GT points are resampled from the official mesh rather than taken from a point-cloud archive. GSO CLIP uses 12 views at 30-degree azimuth intervals and 20-degree elevation with the main-table CLIP evaluator. Uni3D uses the main-table point–text scorer; GSO ULIP-2 uses 10,000 colored surface points and the frozen PointBERT/text encoders. All semantic similarities are computed without GT-selected rotation. Scores should be compared within the GSO panel, rather than numerically equated to other datasets.

Paired interpretation. On the 999 valid BASE/***DreamBridge*** pairs, ***DreamBridge*** has higher NC-24 and ULIP-2 point estimates. The paired 95% interval for the CD-24 difference includes zero; CLIP and Uni3D are descriptive supplementary endpoints. The aggregate results on the fixed 1,000-asset evaluation set are descriptive. The evaluation does not establish the absence of pretraining overlap.

D.6 Text–3D Semantic Metrics

We evaluate semantic alignment between the input text prompt and the generated 3D asset using CLIP, Uni3D, and ULIP-2. CLIP measures semantic consistency through multi-view 2D renderings, whereas Uni3D and ULIP-2 directly compare text and 3D representations in their respective joint embedding spaces.

- *CLIP (Radford et al., 2021).* For each rendered view R_v of a generated asset and its original text prompt T, we compute the cosine similarity between the CLIP image and text embeddings:

$$s_{\mathrm{CLIP}}(T, R_v) = \frac{f_{\mathrm{text}}(T)^\top f_{\mathrm{img}}(R_v)}{\|f_{\mathrm{text}}(T)\|_2 \, \|f_{\mathrm{img}}(R_v)\|_2}.$$

  The multi-view CLIP score is then computed as

$$\text{CLIP-}V = \frac{100}{V} \sum_{v=1}^{V} s_{\mathrm{CLIP}}(T, R_v),$$

  where V denotes the number of rendered views. The main CLIP evaluator uses OpenAI ViT-B/32. The 12-view evaluation on BridgeStruct-1.5K and the eight-view evaluation on Toys4K therefore follow different rendering protocols despite sharing the same score scale. Any results obtained with ViT-L/14 are labeled separately.

- *Uni3D Zhou et al. (2024) and ULIP-2 Xue et al. (2024).* Uni3D and ULIP-2 measure semantic consistency directly between the input text and a point-based representation of the generated 3D asset. For a point cloud $P = \{\mathbf{p}_i\}_{i=1}^{M}$, point coordinates are centered at the sample centroid and normalized to unit maximum radius:

$$\widetilde{\mathbf{p}}_i = \frac{\mathbf{p}_i - \bar{\mathbf{p}}}{\max_j \|\mathbf{p}_j - \bar{\mathbf{p}}\|_2}, \qquad \bar{\mathbf{p}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{p}_i.$$

  Let $f^m_{\text{text}}(T)$ and $f^m_{\text{3D}}(P)$ denote the text and 3D embeddings produced by evaluator m, where $m \in \{\text{Uni3D}, \text{ULIP-2}\}$. The corresponding text–3D similarity is computed as

$$s_m(T, P) = \frac{f^m_{\text{text}}(T)^\top f^m_{\text{3D}}(P)}{\|f^m_{\text{text}}(T)\|_2 \, \|f^m_{\text{3D}}(P)\|_2}.$$

  Both Uni3D and ULIP-2 scores are reported as unscaled cosine similarities.

  For the native BASE/***DreamBridge*** evaluators on BridgeStruct-1.5K and Toys4K, ULIP-2 uses PointBERT with 8,192 surface-sampled XYZ points and its aligned text encoder. Uni3D uses 10,000 surface-sampled points with XYZ plus constant RGB channels set to $100/255$; sampled mesh vertex or UV colors are not used by these two BASE/***DreamBridge*** point scorers. The point clouds are centered at their sample centroid and scaled to unit maximum radius. Both text scorers use the frozen `SimpleTokenizer` with a 77-token context, including the start and end tokens; longer token sequences are truncated to 77. The Uni3D text encoder is EVA02-E-14-plus. Invalid, nonfinite, or degenerate mesh point clouds receive an explicit unavailable value rather than a zero similarity. On Toys4K, LGM's separately evaluated point proxy uses Gaussian centers with spherical-harmonic DC colors, while the Hunyuan shape-only configuration uses fixed gray appearance. LGM point-semantic entries on BridgeStruct-1.5K are unavailable under this evaluation protocol, and no Toys4K proxy score is substituted into the corresponding BridgeStruct-1.5K entries. The GT-selected rotation used for geometric evaluation is not applied to these semantic metrics. Saved ULIP-2 values originally stored on a cosine-times-100 scale are converted back to unscaled cosine similarities for the reported tables.

PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) are secondary measures of appearance agreement with Teacher-IT renderings at matched viewpoints.

D.7 Validity, Coverage, and Paired Statistics

On BridgeStruct-1.5K and Toys4K, metrics are aggregated at the prompt level. A prompt-level value is valid only when all constituent values required by its fixed evaluation record are valid; missing constituents are not averaged away. Each BridgeStruct and Toys4K table mean uses that method's valid prompt set for the specified metric. The GSO BASE/***DreamBridge*** cells instead use the common 999 valid assets for all five reported metrics. Generation coverage, metric coverage, and paired coverage are distinct quantities. Invalid outputs and unavailable metrics remain explicit in the coverage accounting and are not replaced with zero, a worst-case sentinel, or an output from another method.

For metric q, define $\mathcal{V}_{m,q}$ as the valid prompt set of method m. The marginal mean is

$$\overline{q}_m = \frac{1}{|\mathcal{V}_{m,q}|} \sum_{i \in \mathcal{V}_{m,q}} q_{m,i}. \tag{16}$$

For the comparison with BASE, paired analysis uses $\mathcal{P}_q = \mathcal{V}_{\text{ours},q} \cap \mathcal{V}_{\text{base},q}$ and differences $d_i = q_{\text{ours},i} - q_{\text{base},i}$. Consequently, the mean paired difference need not equal the difference between the two marginal table means. Positive differences favor ***DreamBridge*** for higher-is-better metrics; the interpretation is reversed for error metrics.

The split analysis resamples paired prompts with replacement within each split, using 10,000 bootstrap samples. The 2.5th and 97.5th percentiles of the bootstrap mean differences form pointwise 95% intervals. These intervals are not simultaneous across splits or metrics and do not, by themselves, establish that one split benefits more than another.

On Toys4K, LN3Diff and 3DTopia-XL have 1,247 and 1,248 valid generated assets, respectively; the other evaluated methods have 1,250. A metric that is inapplicable remains unavailable even when

Table 10: Verified prompt coverage for the main paired BridgeStruct-1.5K analysis. The target population is 1,536 prompts. Metric-specific validity, rather than a single common denominator, determines each mean.

Metric	BASE valid	***DreamBridge*** valid	Paired valid
SS IoU and SS F1	1,527	1,525	1,518
CLIP-12	1,528	1,526	1,519
Uni3D	1,528	1,526	1,519
ULIP-2	1,528	1,526	1,519

a generation exists. On GSO, BASE has 1,000 valid generations and ***DreamBridge*** has 999; the latter's G471 output retains an engineering failure. The BASE/***DreamBridge*** GSO cells use the 999 valid pairs. External baselines are reported separately under their method-specific protocols. Paired GSO uncertainty uses asset-level bootstrap resampling; a separate identical-prompt-cluster sensitivity analysis addresses repeated titles. Neither analysis proves that these assets were unseen in all model pretraining corpora.

E ADDITIONAL RESULTS AND QUALITATIVE ANALYSIS

E.1 UNALIGNED TOYS4K GEOMETRY

Table 11 retains the original-orientation evaluation alongside the aligned geometry reported in the main text. Its reference is the GT mesh, while the main BridgeStruct-1.5K structural comparisons are teacher-referenced. These protocols address different questions and their absolute scores should not be compared across datasets.

Table 11: Toys4K geometry in the original orientation. LGM uses the Gaussian proxy described in Appendix D.4. Hunyuan3D-2 is the shape-only configuration. Teacher-II is an unranked reference.

Method	Valid	CD-raw ↓	NC-raw ↑	Sil. IoU ↑
LGM	1,250	0.0313	0.5146	–
InstantMesh	1,250	0.0606	0.5158	0.1968
Direct2.5	1,250	0.0599	0.5872	0.4285
Hunyuan3D-2	1,250	0.0557	0.4795	0.1904
Teacher-II	1,250	0.0580	0.5031	0.3673
Shap-E	1,250	0.0583	0.5087	0.2123
LN3Diff	1,247	0.0500	0.5660	0.4342
3DTopia-XL	1,248	0.0403	0.5978	0.3110
BASE	1,250	0.0570	0.5028	0.3811
DreamBridge	1,250	0.0559	0.5086	0.3823

E.2 ANCHOR REPLAY AND REPLAY-FRACTION COMPARISONS

Replay versus teacher-only adaptation. We reuse an earlier 2,048-prompt development experiment comparing teacher-only adaptation (R0) with 90% teacher targets and 10% frozen-base replay (R10). Both variants use identical initialization, rank-16 LoRA, the same 5,000-update schedule, and the same 96-prompt validation protocol. Neither variant uses an additional semantic-direction loss. This development protocol is separate from the formal 8K benchmark evaluation.

The direction diagnostic compares conditional-minus-unconditional velocities to those of the frozen base model at the same latent state, time, and text condition:

$$d_\phi = v_{\theta_0,\phi}(x,t,c(p)) - v_{\theta_0,\phi}(x,t,c(\varnothing)), \qquad q_{\mathrm{CFG}} = \cos(d_\phi, d_0),$$

where d_0 uses the frozen base model. This diagnostic measures retention of the pretrained text-conditioned response. The table reports descriptive validation means rather than the CLIP-12 metric of the main benchmark.

Table 12: Archived teacher-only and 10% replay development comparisons at matched updates. Neither arm uses an additional semantic-direction loss. The higher value in each checkpoint/metric pair is bold.

	SS F1↑		CFG cosine↑	
Updates	R0	R10	R0	R10
3,000	**0.162025**	0.161960	0.5570	**0.5974**
4,000	0.1587	**0.1612**	0.5453	**0.5818**
5,000	0.1580	**0.1602**	0.5431	**0.5785**

Table 13: Replay-fraction comparison in the earlier 2K development protocol. Entries are changes from BASE at a common 3,000-step checkpoint. Chamfer uses voxel units; CLIP-4 changes are multiplied by 100.

Replay	ΔSS IoU↑	ΔSS CD↓	ΔCLIP-4↑
10%	**+0.03627**	**-1.22901**	-0.724
25%	+0.02999	-0.98055	**+0.225**
40%	+0.02473	-0.82317	+0.005

Choosing the replay fraction. A separate earlier 2K development experiment compares R10, R25, and R40 from identical LoRA initialization at a common 3,000-step checkpoint. SS IoU and coordinate Chamfer use 64 validation prompts with two generated samples per prompt; CLIP-4 uses a separate 16-prompt subset with one generated sample per prompt. Table 13 reports changes from BASE. R10 yields the largest structural improvements, motivating our choice of $\rho = 0.1$ under structure-prioritized validation.

E.3 Presentation-Suffix Ablation

This independent development experiment compares 2K variants trained with and without the fixed FLUX presentation suffix under matched training and evaluation settings. We evaluate on 256 prompts and report the downstream semantic metrics on 252 complete prompt pairs.

Table 14 reports CLIP-12 and Uni3D. Using the suffix improves CLIP-12 from 30.06 to 30.17 and Uni3D from 0.1955 to 0.1986. The paired CLIP-12 difference is $+0.114$ on the $\times 100$ scale (95% CI $[0.0003, 0.227]$), and the Uni3D difference is $+0.00308$ (95% CI $[0.00081, 0.00543]$). These results indicate a modest improvement in the reported downstream semantic-alignment metrics.

E.4 Additional Qualitative Comparisons

Qualitative comparisons on BridgeStruct-1.5K. Figure 5 presents additional prompt-matched comparisons on BridgeStruct-1.5K, complementing the qualitative results in the main text. The examples enable visual comparison of overall shape, part arrangement, and geometric detail across methods. Where available, normal and color renderings are shown from the same viewpoint; LGM+MVDream is shown with color renderings only.

Qualitative comparisons on Toys4K. Figure 6 presents qualitative comparisons on Toys4K, complementing the quantitative evaluation in Table 1. These examples illustrate differences in overall shape, part arrangement, and geometric detail across the compared methods.

E.5 User Study

Automatic evaluation metrics for text-to-3D generation primarily assess semantic consistency between generated 3D assets and input text prompts, but may not fully capture whether fine-grained object attributes, structural components, and visual details specified in the text are faithfully reflected in the generated results. To complement the quantitative evaluation, we conduct a pairwise human preference study comparing 3D assets generated by the baseline method and its variant integrated

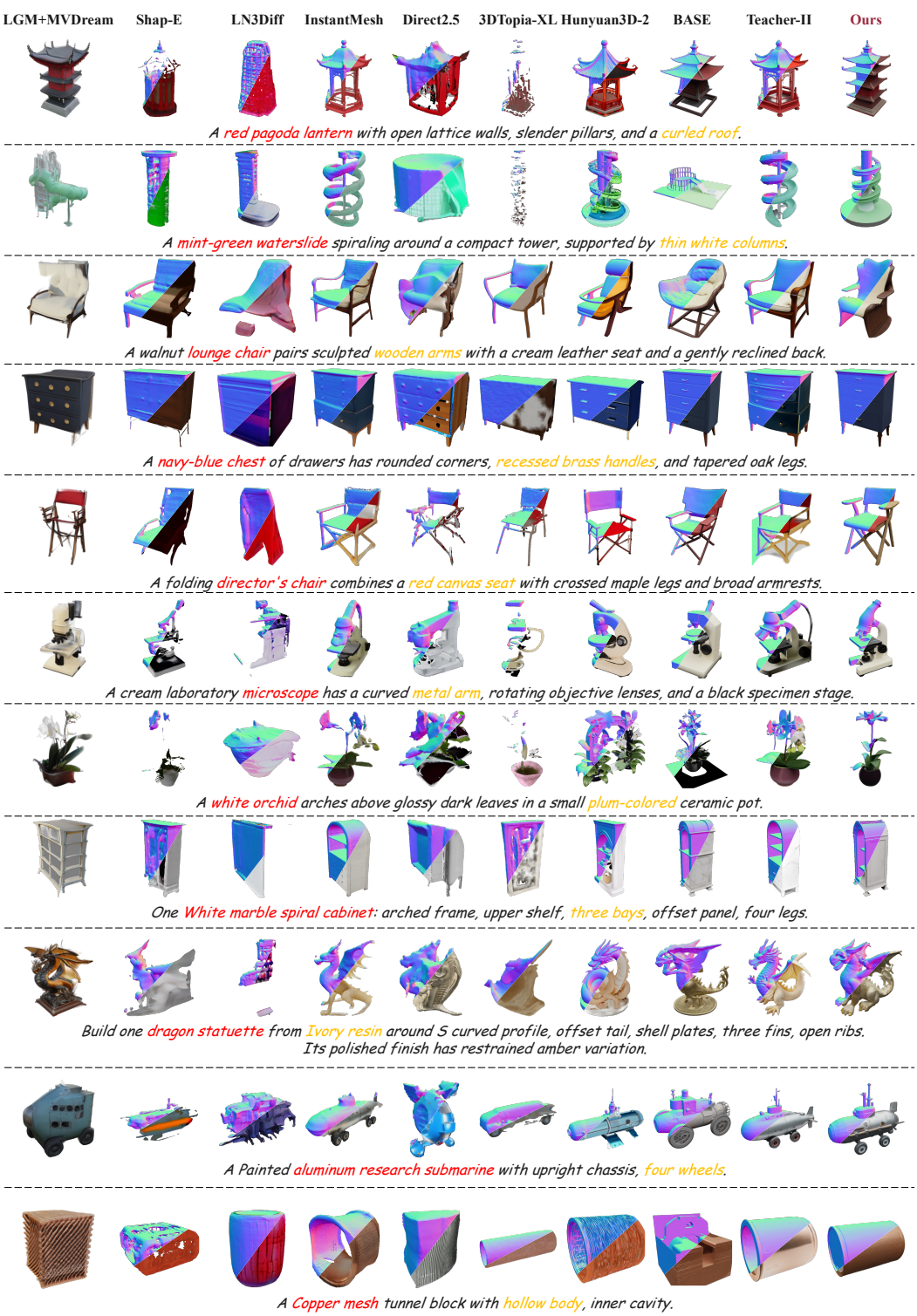

Figure 5: **Additional qualitative comparisons.** We compare ***DreamBridge*** with Shap-E, LGM+MVDream, LN3Diff, InstantMesh, Direct2.5, 3DTopia-XL, Hunyuan3D-2, TRELLIS (BASE), and Teacher-II across diverse text prompts. These examples further illustrate ***DreamBridge***'s ability to produce coherent geometry with distinct structural components and open regions. For methods with available mesh normals, each result is diagonally divided into a normal rendering in the upper-left half and a color rendering in the lower-right half, both shown from the same viewpoint. LGM+MVDream is shown with color renderings only, as mesh normal renderings are unavailable in our pipeline.

Table 14: Presentation-suffix Ablation.

2K variant	CLIP-12↑	Uni3D↑
w/o presentation suffix	30.06	0.1955
With presentation suffix	**30.17**	**0.1986**

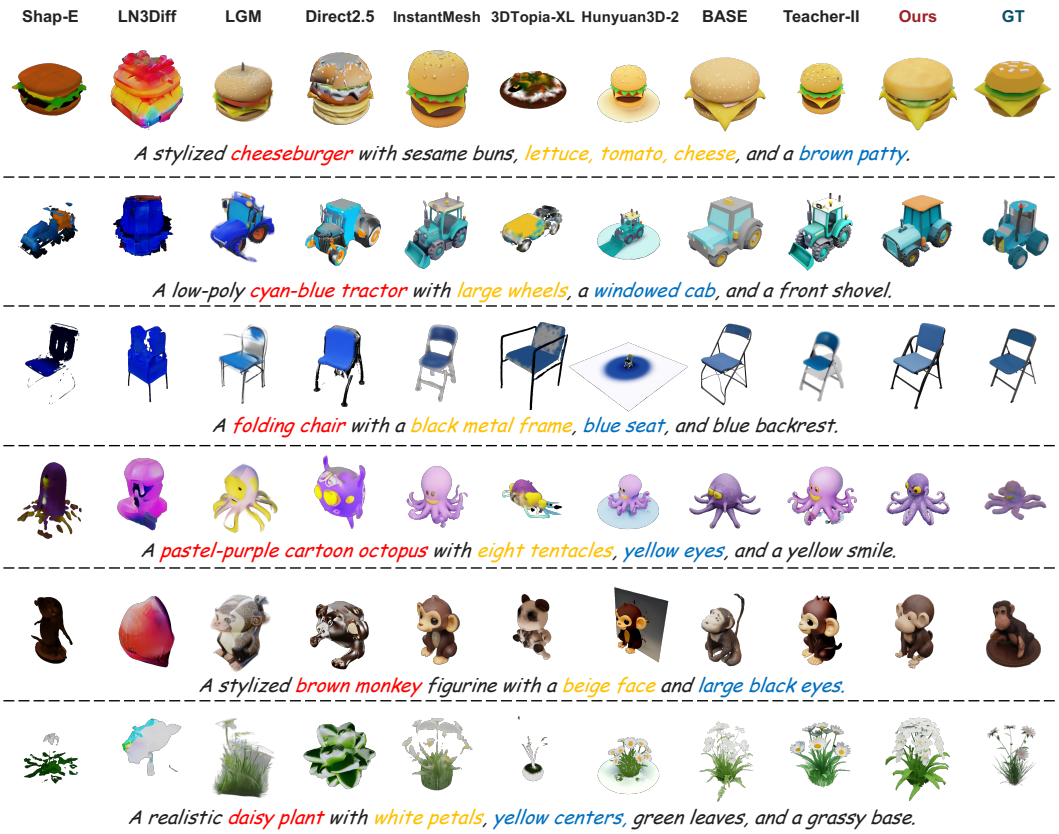

Figure 6: **Qualitative comparisons on Toys4K.** Generated assets are shown for the corresponding text prompts to compare overall shape, part arrangement, and geometric detail.

with ***DreamBridge***. For each comparison, participants are presented with the corresponding textual description together with multi-view renderings of two generated 3D assets and are asked to select the result that more faithfully matches the provided description.

Stimuli Preparation. We construct a fixed set of 20 pairwise comparison tasks from BridgeStruct-1.5K. Each task is assigned to a predefined category, and its prompt is sampled from that category before the study. For each prompt, we pair the 3D assets generated by the baseline method and ***DreamBridge*** under the same text condition. The two assets are rendered from four predefined viewpoints using an identical rendering configuration, including the same camera settings, object scale, lighting, and background. The target object category, a concise textual description, and the full generation prompt are presented together with the paired multi-view renderings.

Evaluation Protocol. The primary user-study analysis includes 22 participants who completed all 20 pairwise comparisons. We adopt a two-alternative forced-choice (2AFC) protocol Zhang et al. (2018), with every participant evaluating the same fixed set of 20 comparison tasks. For each task, the corresponding textual information is displayed above two sets of four-view renderings without revealing the underlying generation methods. The assignment of the baseline and ***DreamBridge*** outputs to A (Left) and B (Right) is randomized to mitigate positional bias.

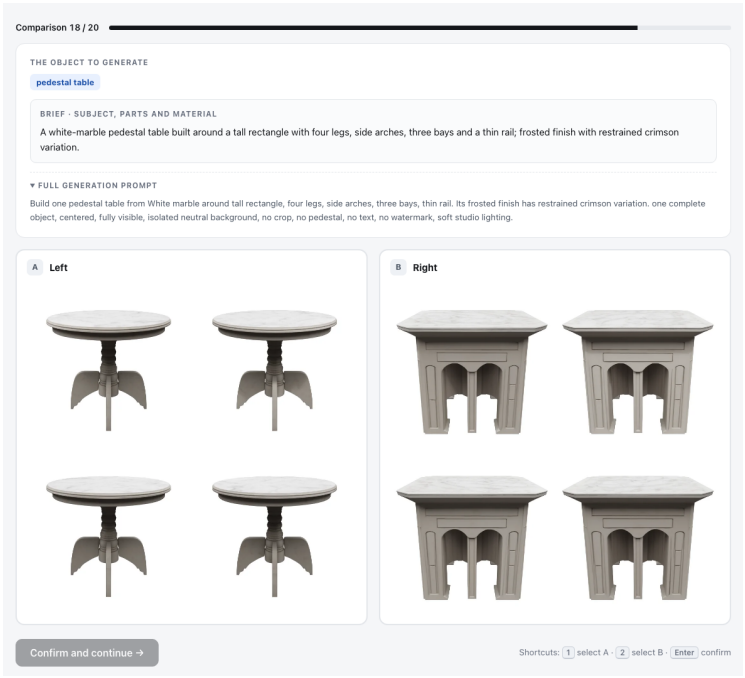

Figure 7: **User study interface.** For each comparison, participants are presented with the target object category, a concise textual description, and the full generation prompt, together with four-view renderings of two anonymized candidate 3D assets labeled A (Left) and B (Right). Participants select the candidate that better matches the textual description and confirm their choice before proceeding to the next comparison.

Participants are asked to determine which of the two candidate 3D assets better matches the provided textual description. They are instructed to select the candidate that more faithfully reflects the specified object identity, geometry, structural components, materials, appearance, and other explicitly described visual attributes. As shown in Fig. 7, participants select one anonymized candidate and confirm their choice before proceeding to the next comparison.

User Study Results. Across the 22 completed participant sessions and 20 comparison tasks, we collect 440 valid pairwise judgments. As shown in Fig. 8, ***DreamBridge*** is preferred in 349 comparisons, corresponding to an overall preference rate of 79.3%, compared with 20.7% for the baseline. At the category level, ***DreamBridge*** receives a majority of preferences in seven of the eight prompt categories, with preference rates ranging from 68.2% to 100.0% among these categories. These results indicate that the overall preference is observed across most evaluated prompt categories rather than being concentrated in only a small subset of comparison cases.

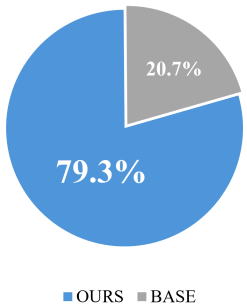

Figure 8: **User study results.**

F Reproducibility Identifiers

The formal adaptation uses the frozen code and model revisions in Table 15. The accepted training manifest is the frozen 8,062-row v3 file; the original 8,192-prompt universe and the 130 failed-image prompts remain separately auditable. Validation and six test-split memberships are fixed independently of checkpoint selection. The three formal training seeds are 2027, 2026, and 2028, with a common selected update 12,000; the smaller-data control uses update 3,000.

Table 15: Pinned code and pretrained-model revisions for formal adaptation.

Component	Revision
TRELLIS code	`442aa1e1afb9014e80681d3bf604e8d728a86ee7`
TRELLIS-text-xlarge	`e0b00432b8e3a8ecee0df806ab1df9f7281f2be4`
TRELLIS-image-large	`25e0d31ffbebe4b5a97464dd851910efc3002d96`
FLUX.1-dev	`3de623fc3c33e44ffbe2bad470d0f45bccf2eb21`

G ALGORITHM

Algorithm 1 summarizes the three training components of ***DreamBridge*** and its text-only generation path. Here $\mathcal{A}_G$ applies the frozen 2D acceptance gates, selects the highest ranked valid image in the initial candidate pool, and uses the first valid slot in each later recovery round if the earlier pools failed; it returns $\varnothing$ if all candidates fail. For $b \in \{T, R\}$, let $(M_T, q_T) = (\mathcal{T}_{\mathrm{SS}}, I_p)$ and $(M_R, q_R) = (\mathcal{B}_{\mathrm{SS}}, p)$. K_p^b contains the prescribed sample indices with source-specific noise, while V_b applies that source's latent and decoded-occupancy checks. $\mathrm{NextBatch}$ follows the fixed category/length/complexity schedule, cycles valid cached samples, and consumes the shuffled source slots without substituting one source for the other. The image generator and image-conditioned teacher are used only in offline cache construction.

We use $D = 8 \cdot 16^3$ and $\sigma_{\min} = 10^{-5}$. The allocation in Algorithm 1 yields exactly 288 teacher and 32 replay positions over each ten global batches of 32; each batch loss is the uniform mean of its element-normalized sample losses, with no further source weight. The frozen text encoder supplies either the original or empty-prompt condition. $\mathrm{TextSSSample}$ denotes the original text-conditioned SS sampler with the adapted velocity field and classifier-free guidance; the SS decoder, text-conditioned SLat sampler, and 3D decoders retain their frozen inference configuration. The exact sampling settings are described in Sec. B.4.

Algorithm 1 ***DreamBridge***: offline targets, structural adaptation, and text-only inference

Require: Prompts $\mathcal{P}$; suffix s; frozen generators, encoders, and decoders; updates U; query $p^\star$
Ensure: Trained LoRA parameters ϕ and generated 3D asset for $p^\star$

Offline teacher and replay caches

$\mathcal{C}_T, \mathcal{C}_R \leftarrow \varnothing$
for all $p \in \mathcal{P}$ **do**
  $I_p \leftarrow \mathcal{A}_G(p \oplus s)$
  if $I_p \neq \varnothing$ **then**
    for all $b \in \{T, R\}$ **do**
      for all $k \in K_p^b$ **do**
        $z^b_{0,p,k} \leftarrow M_b(q_b; \xi^b_{p,k})$
        if $V_b(z^b_{0,p,k}) = 1$ **then**
          $\mathcal{C}_b \leftarrow \mathcal{C}_b \cup \{(p, z^b_{0,p,k})\}$
        end if
      end for
    end for
  end if
end for

Native latent distillation with anchor replay

for $u = 1, \ldots, U$ **do**
  if $u \bmod 10 = 1$ **then**
    Shuffle 288 teacher slots and 32 replay slots into S
  end if
  $\mathcal{B}_u \leftarrow \mathrm{NextBatch}_{32}(S, \mathcal{C}_T, \mathcal{C}_R)$
  $\mathcal{L}_u \leftarrow 0$
  for all $(p, z_0) \in \mathcal{B}_u$ **do**
    $a \sim \mathcal{N}(1, 1)$; $t \leftarrow \mathrm{sigmoid}(a)$
    $\epsilon \sim \mathcal{N}(0, I)$; $m \sim \mathrm{Bernoulli}(0.1)$
    $x_t \leftarrow (1-t)z_0 + [\sigma_{\min} + (1-\sigma_{\min})t]\epsilon$
    $u_t \leftarrow (1-\sigma_{\min})\epsilon - z_0$
    $\widetilde{c} \leftarrow c(\varnothing)$ if $m = 1$, else $c(p)$
    $\widehat{u} \leftarrow f_{\theta_0,\phi}(x_t, 1000t, \widetilde{c})$
    $\mathcal{L}_u \leftarrow \mathcal{L}_u + \|\widehat{u} - u_t\|_2^2/(32D)$
  end for
  Update only ϕ with AdamW using $\nabla_\phi \mathcal{L}_u$
end for

Text-only inference

$z_0^\star \leftarrow \mathrm{TextSSSample}(p^\star; f_{\theta_0,\phi})$
$C^\star \leftarrow \mathrm{Dec}_{\mathrm{SS}}(z_0^\star)$
$h^\star \leftarrow \mathrm{TextSLatSample}(p^\star, C^\star)$
return $\mathrm{Dec}_{\mathrm{3D}}(h^\star)$