BACKDROP: ENABLING ANY-REFERENCE 3D GENERATION WITH ONE LINE OF CODE

Anonymous authors
Paper under double-blind review

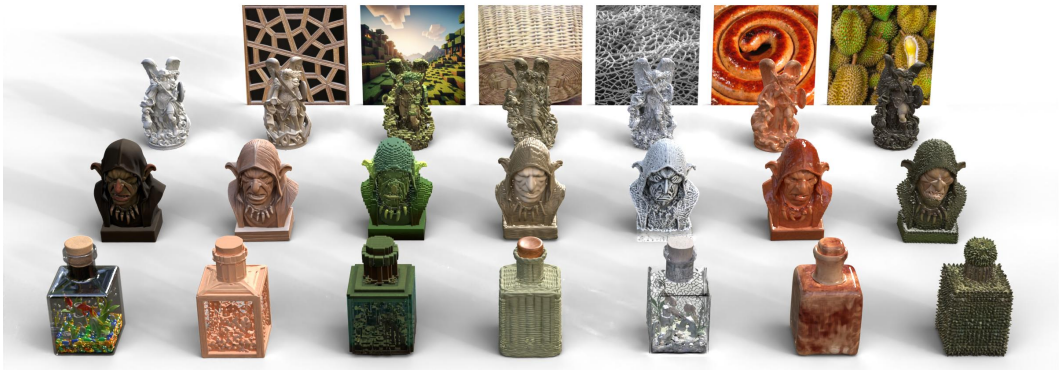


Figure 1: BACKDROP is a one-line-of-code modification that plugs into existing 3D stylization methods and generators, enabling controllable 3D generation from arbitrary image references. It transfers both appearance and geometric style while preserving coherent object structure.

ABSTRACT

We propose BACKDROP, a one-line-of-code intervention that enables controllable 3D generation from arbitrary image references. Existing image-conditioned 3D generators provide a convenient interface for reference-based control, but this interface implicitly assumes object-centric images. When the reference is instead a texture, scene, painting, or material photograph, existing stylization methods can severely distort or fragment the generated geometry. We trace this failure to a mismatch in conditioning structure: object-centric references contain a large population of background patch tokens that absorb substantial cross-attention in deeper generator blocks, while full-frame references do not. Removing these tokens from object references reproduces the failure, while restoring them to non-object references repairs it. BACKDROP exploits this observation by appending a cached set of patch tokens from a single blank image to the reference condition. This restores the generator's expected fallback population without changing the reference, retraining the model, optimizing per instance, or altering the underlying control method. Across diverse textures, scenes and artistic images, and with different 3D stylization methods, BACKDROP consistently improves geometric integrity while preserving strong reference transfer and content shape, making arbitrary images reliable controls for 3D generation. Results and code are available at https://backdrop-3d.pages.dev/.

1 INTRODUCTION

The visual identity of a game or animated film is often established through concept art and mood boards before its 3D assets exist. Depending on the reference, an object may need to become blocky, perforated, greasy, woven, or otherwise shaped by a visual language that affects both appearance and geometry. Image-to-3D generators such as TRELLIS.2 can already turn a single object image into a textured asset in seconds (Xiang et al., 2025; 2026). A natural next step is to let artists guide this process with the references they already use, including material photographs, paintings, concept art, and scene screenshots, rather than requiring another isolated object.

Style-controlled 3D generation provides a natural interface for this goal: a content image specifies *what* to generate, while a reference image specifies *how it should look and be shaped*. Methods such as StyleSculptor and MorphAny3D realize this idea by injecting a second image into a frozen 3D

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.

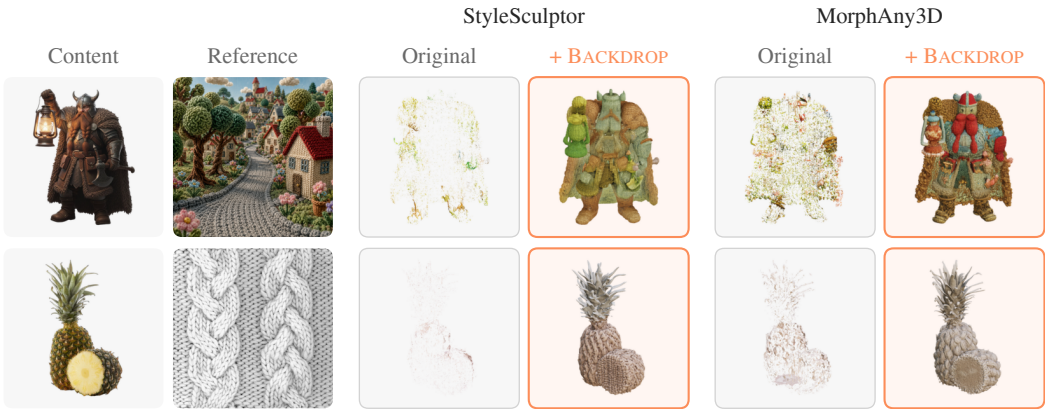


Figure 2: Non-object references can cause geometric collapse in existing 3D stylization methods. For the two content–reference pairs shown, StyleSculptor and MorphAny3D on TRELLIS.2 severely erode or fragment the content geometry. Adding BACKDROP preserves recognizable structure while transferring the reference's material and appearance.

generator (Qu et al., 2025; Sun et al., 2026). Yet this interface is much less general than it appears. On TRELLIS.2, object references often work, while full-frame textures, scenes, and paintings can severely erode the content geometry or reduce it to fragments (Figure 2). The failure is especially striking because these references are perfectly valid images and often contain exactly the visual structure an artist wants to transfer.

We trace this failure to a seemingly irrelevant part of the conditioning image: its background. TRELLIS.2 is designed around object-centric inputs, where the object occupies only part of the frame and the remaining patches depict empty background. These background patches are not ignored. In the deeper shape-transformer blocks we examine, they absorb most of the cross-attention mass (more analysis in Figure 3), effectively providing a fallback population when a 3D query does not need strong visual evidence from the object itself. Full-frame references remove this fallback. Attention that would normally rest on background patches is instead forced onto actual style patches, causing the reference to be applied too aggressively and destabilizing geometry. A set of causal interventions (Section 3.1) supports this interpretation. Removing background patches from an object reference reproduces the failure, while restoring a background-like token population to a scene or texture repairs it. Adding the same population to an object reference has little effect. These results suggest that the problem is not that non-object references lack useful style information, but that they violate the conditioning structure the frozen generator expects.

Based on this observation, we introduce BACKDROP, a one-line code augmentation that makes existing stylization methods compatible with arbitrary image references (Figure 1). We encode an all-black image once and append its cached patch tokens to the reference conditioning in the shape and texture stages. These tokens restore the missing fallback population without modifying the reference image or changing how the underlying stylization method combines content and style. The same cached tokens are reused for every reference, requiring no training, per-reference optimization, or additional encoder evaluation, with only minimal computational overhead.

We evaluate style transfer, structural integrity, and shape preservation as separate objectives. Our benchmark pairs diverse content objects with references spanning material textures, scenes, and artistic images, and compares against state-of-the-art 3D stylization methods and stylize-then-generate pipelines. Across these settings, BACKDROP achieves the highest style similarity among the evaluated methods while preserving intact geometry and content shape. Our contributions are:

- We identify missing background conditioning as a cause of geometric collapse under non-object references, revealing a hidden dependence inherited from the frozen 3D generator.
- We introduce BACKDROP, a one-line code conditioning augmentation that restores this missing fallback with a single cached token sequence, enabling textures, scenes, and artistic images to serve as references without training or per-reference optimization with very little additional cost.
- We introduce an evaluation benchmark spanning diverse reference types and separately measure style transfer, structural integrity, and shape preservation, supported by causal interventions and cross-backbone comparisons.

2 Related Work

3D stylization. Many 3D stylization methods transfer appearance while preserving geometry, through per-scene radiance-field optimization (Zhang et al., 2022), zero-shot feature transformations (Liu et al., 2023), or feed-forward stylized reconstruction (Wang et al., 2025; Liu et al., 2026). Non-object references are already useful in this setting: scene photographs can guide texture and lighting (Zhou et al., 2024), while IPDreamer uses image prompts for appearance control in text-to-3D optimization and notes the difficulty of references without a clear subject (Zeng et al., 2025). Geometric stylization further modifies shape by deforming meshes toward text or image references (Dinh et al., 2025; Choi et al., 2026), or optimizing radiance-field and Gaussian geometry (Jung et al., 2024; Bang et al., 2026). These approaches typically optimize each asset or scene. We instead study style-controlled generation, where a content image and a style reference jointly guide geometry and appearance without requiring a source 3D asset.

Style control in image-to-3D generators. Image-to-3D generators condition on image tokens through cross-attention (Zhang et al., 2024; Wu et al., 2024; Xiang et al., 2025; 2026), providing a natural interface for style control, analogous to attention sharing and feature injection in 2D diffusion models (Hertz et al., 2024; Chung et al., 2024; Alaluf et al., 2024). StyleSculptor injects reference features through attention and separates geometric from textural control (Qu et al., 2025). MorphAny3D blends cross-attention outputs from two images (Sun et al., 2026), while Fuse3D concatenates conditioning tokens from selected image regions (Jin et al., 2025). Style control can also be learned by fine-tuning (Wang et al., 2026) or applied before generation with 2D stylizers (Gao et al., 2026; Zhu et al., 2026). Our focus is the reference conditioning supplied to an existing stylization method. Similar to feature masking for reducing content leakage in 2D stylization (Zhu et al., 2025), BACKDROP modifies the conditioning itself while leaving the method's content-style combination rule unchanged.

What generators read from the conditioning image. Many image-to-3D generators are trained on isolated-object renders (Deitke et al., 2023b;a), often with the subject placed against a uniform background. Cue3D shows that such generators rely strongly on object silhouette (Li et al., 2025), while other work adapts conditioning to handle occlusion (Wu et al., 2025b) or surrounding scene context (Chen et al., 2026). Background handling also differs across models: Direct3D-S2 explicitly restricts cross-attention to foreground tokens (Wu et al., 2025a). We instead study a frozen generator trained with background tokens present. Concurrent work UMI3D reports that TRELLIS-family models use global tokens at 3D locations lacking direct visual support (Qu et al., 2026). Our measurements refine this picture across network depth: in TRELLIS.2, global tokens dominate early blocks, while background tokens receive most attention in deeper blocks (Figure 3).

Attention sinks and appended tokens. Transformers can place substantial attention on tokens with little semantic content, including initial tokens in language models (Xiao et al., 2024), background patches in vision transformers (Bondarenko et al., 2023), and start-of-text tokens in text-to-image cross-attention (Ren et al., 2024; Marcos-Manchón et al., 2024). Such tokens can act as sinks or key biases (Gu et al., 2025), although their importance varies across architectures (Kang et al., 2025; Wu & Summa, 2026; Xiao et al., 2024). Related work also introduces extra tokens to redirect attention: registers can be added during training (Darcet et al., 2024) or at test time (Jiang et al., 2025), and CODA appends text-padding embeddings as cross-attention registers in an object-centric diffusion model (Nguyen et al., 2026). Our setting differs in that we restore the background-token population expected by a frozen 3D generator, using tokens from a blank image rather than learned registers.

3 From Missing Background to Any-Reference Control

3.1 Diagnosing the missing background

TRELLIS.2, our primary backbone, generates a 3D asset in three stages (Xiang et al., 2026). A sparse-structure stage first determines the object's coarse voxel support. A shape stage then generates structured latent features (SLat) describing detailed geometry on the active voxels, and a texture stage predicts appearance conditioned on the resulting shape latents. The shape and texture latents are finally decoded into a textured asset.

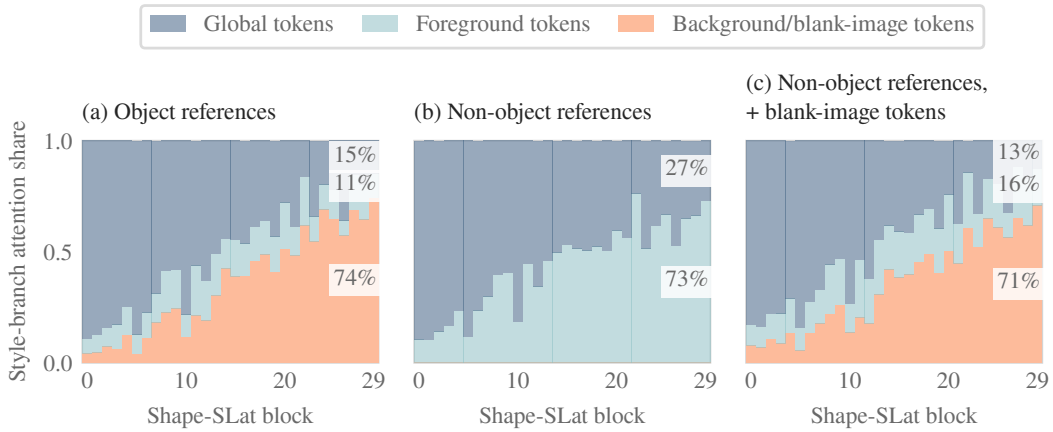


Figure 3: Non-object references shift deep-block attention from background to foreground tokens. We group reference attention into global, foreground, and background tokens across the shape-SLat transformer. Object references devote most deep-block attention to native background tokens, whereas full-frame references have no such population and concentrate attention on their original patches. Appending BACKDROP's blank-image tokens restores a similar attention allocation. Results use attention blending at $\alpha = 0.85$ and average 144 object or 432 non-object pairs.

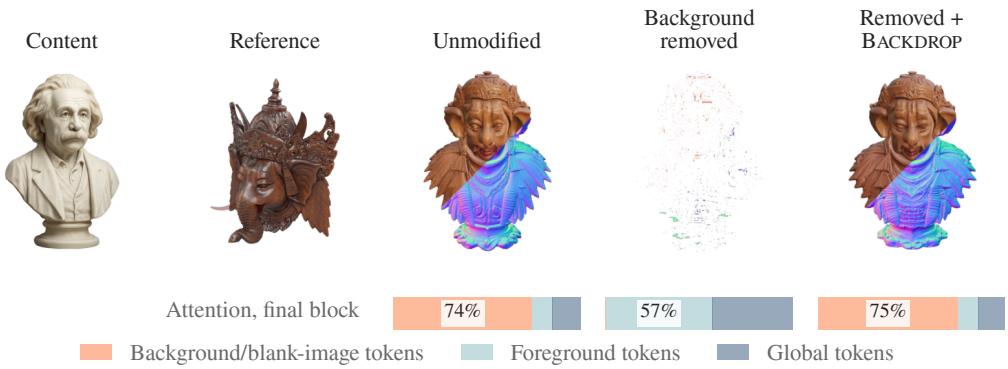


Figure 4: Background tokens are causally linked to geometric integrity. Removing only the background tokens from an object reference shifts most of their attention to foreground tokens and reproduces the collapse seen with non-object references. Replacing them with BACKDROP's blank-image tokens restores both the attention allocation and the geometry. Bars show reference attention in the final shape-SLat block with attention blending at $\alpha = 0.85$, averaged over 144 object pairs.

All three stages receive image conditioning from DINOv3 (Siméoni et al., 2026). The encoder represents an image as a sequence of *patch tokens*, together with *global tokens* consisting of one CLS token and several register tokens. TRELLIS.2 preprocesses object images by placing an isolated subject against a uniform background. We therefore partition its patch tokens into *foreground tokens*, whose source patches intersect the object mask, and *background tokens*, whose source patches lie entirely outside it. Cross-attention then lets each latent query retrieve information from these image tokens.

Reference-guided stylization introduces a second image through this same interface. We first ask how the shape generator uses the different token populations. Figure 3 reveals a strong depth-dependent pattern. For object references, global tokens dominate early blocks, but background tokens absorb most reference attention in deeper blocks. The seemingly empty background is therefore not ignored by the generator.

This population is missing from full-frame references such as scenes, paintings, and textures. Every patch contains visual content, so attention that would otherwise fall on background tokens is redistributed to the reference patches themselves. In the deeper blocks, these patches consequently receive substantially more attention than foreground patches do for object references.

Because object and non-object references differ in many ways beyond background, we isolate this factor directly. For the 144 object-reference pairs in Section 4.1, we remove only the encoded back-

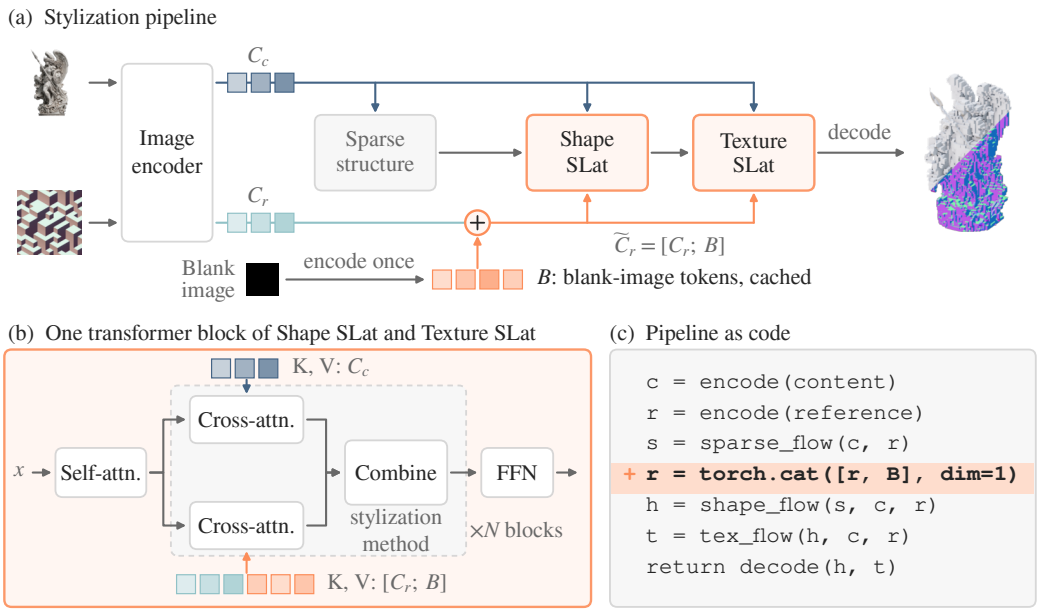


Figure 5: BACKDROP within the TRELLIS.2 stylization pipeline. (a) The image encoder produces content tokens C_c and reference tokens C_r. BACKDROP concatenates C_r with cached blank-image tokens B to form $\widetilde{C}_r = [C_r; B]$, which conditions the shape and texture stages; the sparse-structure stage is unchanged. (b) Each transformer block reads $\widetilde{C}_r$ as keys and values, while the stylization method's original content–reference combination rule is preserved. (c) The full pipeline in pseudocode, with BACKDROP highlighted as a single line. The example uses StyleSculptor with BACKDROP, shown as a textured render and surface normals.

ground tokens while keeping the content image, reference image, style strength, and all remaining tokens fixed. As shown in Figure 4, this single intervention reproduces the non-object failure: attention shifts primarily to foreground tokens and the generated geometry collapses. Replacing the removed tokens with patch tokens encoded from an independent blank image restores both the attention allocation and the geometry.

These interventions identify a hidden dependency in the frozen generator: coherent generation does not require the reference's *own* background, but it does benefit from a background-like token population. Non-object references violate this conditioning structure because every patch carries reference content. BACKDROP restores the missing population.

3.2 BACKDROP: RESTORING THE MISSING BACKGROUND

Given a content image and an arbitrary reference image, our goal is to preserve the content object's identity while transferring the reference's geometric and appearance characteristics. BACKDROP does so with a single operation at the conditioning interface: it appends patch tokens from a blank image to the reference tokens, as illustrated in Figure 5. As the name suggests, it places a backdrop behind the reference at the conditioning interface.

We encode a blank image using the generator's image encoder, discard its global tokens, and retain its patch tokens as a sequence B, which we call the *blank-image tokens*. Given the original reference sequence C_r, we form $\widetilde{C}_r = [C_r; B]$ by concatenating along the token dimension. With the cached blank-image tokens ready for concatenation, the inference-time operation is

```
reference_cond = torch.cat([reference_cond, blank_tokens], dim=1)
```

All original reference tokens are retained, and the same B is reused for every input under a fixed encoder configuration. BACKDROP therefore requires no training, per-reference fitting, or additional image encoding at inference, and does not modify how the underlying stylization method combines content and reference information.

For TRELLIS.2, we use an all-black 512×512 image matching its native background fill and retain all 1,024 patch tokens after the pipeline's conditioning normalization. Although the input pixels are uniform, the encoded patch tokens remain distinct because the image encoder incorporates spatial information. Preserving this token population is important: collapsing it to a single repeated feature does not reproduce the same attention behavior. We apply BACKDROP to the shape and texture SLat stages while leaving the sparse-structure stage unchanged. Appendix A gives the exact insertion points and encoder settings.

The effect of appending B can be understood directly from cross-attention. For any fixed query, adding blank-image tokens leaves the relative attention weights among the original reference tokens unchanged, while reducing their total attention mass by a common factor. The remaining mass is assigned to B, whose contribution varies naturally across attention heads, transformer blocks, and generation steps. Thus, BACKDROP does not alter *which* original reference patches are preferred relative to one another. Instead, it restores an alternative population to which the frozen generator can route attention, recovering the conditioning structure present for object-centric inputs. Appendix C gives the derivation.

Because BACKDROP acts only on the reference token sequence, it composes directly with existing stylization rules. Section 4.2 evaluates it with attention blending, StyleSculptor (Qu et al., 2025), and MorphAny3D (Sun et al., 2026).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and settings. We use TRELLIS.2 as our primary backbone, with 12 sampling steps and a default style weight of $\alpha = 0.85$ or the corresponding setting for each method. We pair 12 content images from TRELLIS and TRELLIS.2 with 36 non-object references, mostly from the Describable Textures Dataset (DTD) (Cimpoi et al., 2014) and StyleBench (Gao et al., 2026), spanning textures, scenes and artistic images. This yields 432 pairs, with seeds matched across methods. For the object-reference analyses in Section 3.1 and Figure 8, we pair the same contents with 12 Objaverse-XL object references (Deitke et al., 2023a) that retain their native backgrounds.

Baselines. We compare StyleSculptor and MorphAny3D (Qu et al., 2025; Sun et al., 2026) as 3D stylization baselines, testing each with and without BACKDROP on TRELLIS.2. We also evaluate attention blending, a simple weighted average of content and reference cross-attention outputs, which serves as the baseline for the interventions and controls. To compare across backbones, we include the released TRELLIS implementations of StyleSculptor and MorphAny3D. We further compare with *stylize-then-generate* pipelines, which apply IP-Adapter-Plus (Ye et al., 2023) or StyleExpert (Zhu et al., 2026) to the content image before passing it to TRELLIS.2 or Hunyuan3D 2.1 (Team Hunyuan3D et al., 2025).

Evaluation metrics. We report three main metrics. (1) Using Contrastive Style Descriptors (CSD) (Somepalli et al., 2024), we measure style similarity as the mean cosine similarity between textured output renders and the reference. (2) Structural Integrity Score (SIS) measures fragmentation, debris and open boundaries: 1 indicates no additional detected defects, while values below 0.5 denote *collapse*. (3) Bidirectional point-to-mesh F-score (F@2%) measures shape preservation within 2% of the base's longest bounding-box dimension. It captures missing geometry but can also penalize intended shape changes. Both geometry metrics compare against an unstylized base matched by content image, backbone and seed. We also include FID (Heusel et al., 2017) between output-render and style-reference distributions as a standard style-transfer metric. We call a difference *clear* when its 95% bootstrap interval over references excludes zero. Full settings and metric definitions are in Appendices A and B.

4.2 MAIN RESULTS

Table 1 shows the effect of adding BACKDROP while holding each stylization method's settings fixed. Without this augmentation, most outputs collapse under non-object references for all three methods on TRELLIS.2. Adding blank-image tokens keeps nearly all outputs intact (Figure 2) and keeps their shapes closer to the unstylized base. CSD improves clearly for attention blending and

Method	Backbone	CSD ↑	SIS ↑	F@2% ↑	FID ↓	Time (s)
Unstylized base	TRELLIS.2	0.106	–	–	370.7	9.7
Attention blending	TRELLIS.2	0.254	0.128	0.9209	419.0	11.1
+ BACKDROP	TRELLIS.2	**0.331**	0.975	0.9925	358.0	12.1
StyleSculptor (SIGGRAPH Asia'25)	TRELLIS.2	0.242	0.173	0.8689	412.2	40.5
+ BACKDROP	TRELLIS.2	0.321	0.972	0.9773	362.4	40.9
MorphAny3D (CVPR'26)	TRELLIS.2	0.255	0.190	0.9715	421.2	223.6
+ BACKDROP	TRELLIS.2	0.278	0.989	**0.9970**	363.4	234.0
StyleSculptor	TRELLIS	0.288	0.991	0.9841	366.6	28.4
MorphAny3D	TRELLIS	0.309	0.992	0.9909	365.3	429.4
IP-Adapter-Plus (arXiv'23) → 3D	TRELLIS.2	0.219	0.989	0.5856	356.6	16.0
StyleExpert (CVPR'26) → 3D	TRELLIS.2	0.129	0.989	0.6507	364.7	127.7
IP-Adapter-Plus → 3D	Hunyuan3D 2.1	0.228	0.990	0.4925	**353.6**	178.8
StyleExpert → 3D	Hunyuan3D 2.1	0.141	**0.994**	0.5630	363.1	300.8

Table 1: Results on 432 non-object content–reference pairs per row. Methods that stylize during generation use style weight 0.85 or its counterpart. Shaded rows add BACKDROP; bold marks the best value in each metric column. Times are mean generation seconds per case.



Figure 6: Stylize-then-generate pipelines on TRELLIS.2 for four content–reference pairs: the content image stylized by IP-Adapter-Plus or StyleExpert and the resulting 3D asset, alongside the unstylized base and attention blending with BACKDROP at $\alpha = 0.85$. Each 3D view is split diagonally between a textured render (upper left) and surface normals (lower right).

StyleSculptor, while MorphAny3D shows no clear change. FID also decreases for all three methods. The released TRELLIS implementations of StyleSculptor and MorphAny3D produce intact meshes without BACKDROP, indicating that the failure depends on the backbone. Section 4.4 examines this difference through each backbone's attention allocation.

The stylize-then-generate baselines also produce intact meshes, but their lower F@2% scores reveal substantially larger departures from the unstylized base than those produced by methods that stylize during generation. Stylization in 2D can alter the object itself, and those changes carry through to the generated geometry (Figure 6). IP-Adapter-Plus redraws the boots as camouflage army boots, gives the bust a new, faceted face and buries the statue's figure under grass. StyleExpert keeps each object's type but redraws details such as the boots' straps and buckles and the bust's face. This illustrates why structural integrity and shape preservation need separate measures. These pipelines also score clearly lower on CSD than all three methods combined with BACKDROP.

To test whether the benefit extends beyond the default style weight, we sweep α using attention blending (Figure 7). With BACKDROP, style similarity rises throughout the sweep while structural

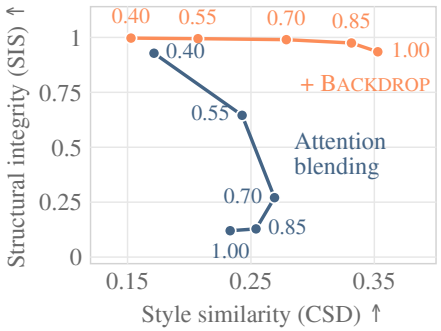


Figure 7: Style similarity (CSD) and structural integrity (SIS) across style weights for attention blending on TRELLIS.2. Point labels show α; each point averages 432 non-object pairs. With BACKDROP, CSD rises throughout the sweep while SIS remains high. Without it, SIS falls as α increases, and CSD declines beyond $\alpha = 0.70$. The points at $\alpha = 0.85$ correspond to the attention-blending rows in Table 1.

Reference conditioning	CSD ↑	SIS ↑
No appended tokens	0.254	0.128
BACKDROP	0.331	0.975
Black border drawn on the image		
Border on a larger canvas	0.314	0.980
Border around a shrunken reference	0.302	0.990
Blank-image token ablations		
Copies of the mean token	0.311	0.391
Blank-image keys, zero values	0.342	0.984

Table 2: Image-border and blank-token ablations with attention blending on TRELLIS.2, evaluated on 432 non-object pairs at $\alpha = 0.85$. Borders add black padding before encoding; BACKDROP and its variants append 1,024 encoded tokens. Mean copies remove token variation, while zero-value tokens preserve keys but remove value contributions. Appendix D reports the full comparisons, F@2%, and paired intervals.

integrity remains high. Without it, CSD peaks at $\alpha = 0.70$, when much of the structure has already been lost, then declines as the geometry collapses further. BACKDROP thus allows stronger stylization while preserving the geometry across the tested weights with very little cost (Table 1).

4.3 ABLATIONS

A natural alternative is to add the background directly to the reference image. We test two black borders: one around the reference on a larger canvas, and one around a shrunken reference on the original canvas (Table 2). Both preserve intact geometry, but yield clearly lower CSD than BACKDROP at the same style weight, with at most a small improvement in SIS. Adding a background in image space therefore works, though separately encoded tokens give a better style result. The constructions also differ within the encoder: self-attention mixes a border with the reference, while BACKDROP encodes the blank image separately and preserves the original reference tokens.

Although the blank image has uniform pixels, DINOv3 incorporates positional information through self-attention (Siméoni et al., 2026), and its output tokens vary with patch position. This variation is already encoded in the features before they reach the generator. Replacing the blank-image tokens with 1,024 copies of their mean preserves the token count and average token, but most outputs collapse, with no clear change in CSD. These mean tokens receive 41.7% of attention in the final shape block, compared with 71.0% for BACKDROP's blank-image tokens over the same 432 pairs. Without their encoded variation, the appended tokens draw less attention and lose most of the structural-integrity gain (Appendix D).

Finally, we separate the blank-image tokens' competition for attention from the values they contribute. We keep their keys but set their projected values to zero, so the tokens receive the same attention share as in BACKDROP but contribute nothing to the attention-weighted sum (Equation 3). This change yields no clear difference in SIS and a small increase in CSD relative to BACKDROP. Zeroing the values is thus a possible variant of BACKDROP. We do not adopt it because it must be implemented inside the cross-attention of each stylization method and backbone, whereas appending tokens changes only the inputs. The result shows that, in this setting, competition from the blank-image keys is sufficient for the gains in structural integrity and style.

4.4 SCOPE AND FAILURES

Our analysis focuses on geometry because its failures are easier to quantify, but the background dependence extends to the texture stage. There too, deep blocks attend mostly to background tokens on object references, and background removal redirects most of that attention to the foreground

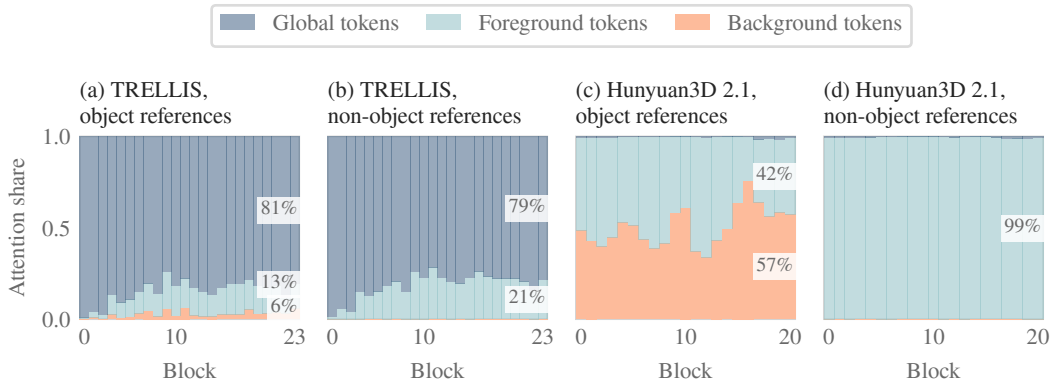


Figure 8: Attention to the reference tokens across depth on two further backbones, with attention blending at $\alpha = 0.85$ and token groups as in Figure 3: (a, b) the 24 structured-latent blocks of TRELLIS and (c, d) the 21 shape-transformer blocks of Hunyuan3D 2.1. Object panels average 144 pairs and non-object panels 432. TRELLIS keeps its deep-block attention on the global tokens whether or not the reference has a background; Hunyuan3D 2.1, like TRELLIS.2, places it on the background and moves it to the foreground on full-frame references.

(Appendix C). Omitting blank-image tokens from this stage produces overexposed and oversaturated textures on otherwise intact shapes in our qualitative comparisons (Appendix F). These observations motivate applying BACKDROP to both shape and texture conditioning.

Across backbones, the diagnosis predicts a need for repair where deep blocks rely on background tokens. We examine this attention allocation on TRELLIS and Hunyuan3D 2.1, both using DINOv2 rather than DINOv3 (Figure 8). TRELLIS assigns most final-block attention to its five global tokens, which are present in every reference, and little to the background; full-frame references leave this allocation largely unchanged. Consistent with the prediction, attention blending produces no collapses on non-object references.

Hunyuan3D 2.1 resembles TRELLIS.2: deep blocks attend mostly to the background, and full-frame references shift this attention to the foreground. However, attention blending introduces a different failure that prevents testing the predicted repair. It generates intact geometry from the reference instead of preserving the content, such as a cube from a texture image, leaving SIS high but F@2% very low (Appendix F). Appending tokens from an all-white image, matching its native background, barely changes this behavior. An architectural difference may explain the result.

An exploratory background-removal test on Hunyuan3D 2.1 uses direct generation without stylization (Appendix F). All 12 content images still produce their objects, but with coarser and mostly inflated shapes. Among the 12 object references, eight remain intact but coarser, three lose structure, and one collapses to its ground plate. The background therefore affects geometry in this setting too, although this is a different task from the stylization test above.

Direct generation also exposes a limit of BACKDROP on TRELLIS.2: restoring a background does not supply the object structure missing from a texture or painting. In stylization, the content image provides this structure, and BACKDROP retains each method's original sparse-structure conditioning. Without a content image, the sparse-structure stage must infer an appropriate coarse shape from the non-object reference alone. Adding blank-image tokens to that stage can help, but the benefit varies across inputs and seeds; no tested setting reliably produces appropriate geometry (Appendix F.1).

5 CONCLUSION

Existing methods for style-controllable 3D generation work well with object references, but on TRELLIS.2 often collapse on textures, scenes, and paintings. We trace this failure to a missing background: deep generator blocks devote much of their attention to the background tokens surrounding isolated objects. BACKDROP repairs this in one line of code by appending cached blank-image tokens to the reference conditioning, with no training, per-reference tuning, or changes to the stylization method. It preserves geometry while improving style transfer across diverse references, allowing textures, scenes, and paintings to serve as reliable style controls for 3D generation.

AI USE STATEMENT

We used large language models to assist with code, analysis scripts, and manuscript drafting. The authors reviewed and verified these contributions and take responsibility for the results, interpretations, and cited sources.

ETHICS STATEMENT

This work involves no human-subject study. Our method inherits biases and limitations of the pretrained models and may facilitate misleading or infringing 3D content. Users should respect reference-image rights, attribute source artwork where appropriate and disclose generated content when its origin could be misleading.

REPRODUCIBILITY STATEMENT

Appendix A details the implementation, model settings and timing protocol; Appendix B specifies the evaluation data, seed rule, metrics and aggregation. We plan to release the code, evaluation manifests and per-pair results to support reproduction of the reported experiments.

REFERENCES

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024.

Min Hyeok Bang, Jun Hyeong Kim, Seung-Wook Kim, and Se-Ho Lee. Geometry-aware style transfer in 3D gaussian splatting. In *European Conference on Computer Vision (ECCV)*, Malmö, Sweden, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75067–75096, 2023.

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

Changwoon Choi, Hyunsoo Lee, Clément Jambon, Yael Vinker, and Young Min Kim. Image-guided geometric stylization of 3D meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19972–19981, 2026.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8795–8805, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, pp. 2632–2652, 2024.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35799–35813, 2023a. Datasets and Benchmarks Track.

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 (CVPR)*, pp. 13142–13153, 2023b.

Nam Anh Dinh, Itai Lang, Hyunwoo Kim, Oded Stein, and Rana Hanocka. Geometry in style: 3D stylization via surface normal deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28456–28467, 2025.

Junyao Gao, Yanan Sun, Yanchen Liu, Yinhao Tang, Yanhong Zeng, Ding Qi, Kai Chen, and Cairong Zhao. StyleShot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1215–1228, 2026.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, pp. 97114–97144, 2025.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4775–4785, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

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

Nick Jiang, Amil Dravid, Alexei A. Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 56557–56595, 2025.

Xuancheng Jin, Rengan Xie, Wenting Zheng, Rui Wang, Hujun Bao, and Yuchi Huo. Fuse3D: Generating 3D assets controlled by multi-image fusion. In *ACM SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12. Association for Computing Machinery, 2025.

Hyunyoung Jung, Seonghyeon Nam, Nikolaos Sarafianos, Sungjoo Yoo, Alexander Sorkine-Hornung, and Rakesh Ranjan. Geometry transfer for stylizing radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8565–8575, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, pp. 87676–87703, 2025.

Xiang Li, Zirui Wang, Zixuan Huang, and James M. Rehg. Cue3D: Quantifying the role of image cues in single-image 3D generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Hanzhou Liu, Jia Huang, Mi Lu, Srikanth Saripalli, and Peng Jiang. Stylos: Multi-view 3D stylization with single-forward Gaussian splatting. In *International Conference on Learning Representations*, 2026.

Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P. Xing. StyleRF: Zero-shot 3D style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8338–8348, 2023.

Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C. SanMiguel, and José M. Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9242–9252, 2024.

Bac Nguyen, Yuhta Takida, Naoki Murata, Chieh-Hsin Lai, Toshimitsu Uesaka, Stefano Ermon, and Yuki Mitsufuji. Improved object-centric diffusion learning with registers and contrastive alignment. In *International Conference on Learning Representations*, pp. 51588–51616, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Zefan Qu, Zhenwei Wang, Haoyuan Wang, Ke Xu, Gerhard Petrus Hancke, and Rynson W. H. Lau. StyleSculptor: Zero-shot style-controllable 3D asset generation with texture-geometry dual guidance. In *ACM SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12. Association for Computing Machinery, 2025.

Zefan Qu, Zhenwei Wang, Gerhard Petrus Hancke, and Rynson W. H. Lau. UMI3D: Robust 3D generation on unconstrained multi-image inputs via simultaneous focus cross-attention routing. *arXiv preprint arXiv:2607.24298*, 2026.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision (ECCV)*, pp. 340–356, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Investigating style similarity in diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 143–160, 2024.

Xiaokun Sun, Zeyu Cai, Hao Tang, Ying Tai, Jian Yang, and Zhenyu Zhang. MorphAny3D: Unleashing the power of structured latent in 3D morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27018–27029, 2026.

Team Hunyuan3D et al. Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material. *arXiv preprint arXiv:2506.15442*, 2025.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. InstantStyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

Kai Wang, Ziheng Ouyang, Xuying Zhang, Ming-Ming Cheng, and Qibin Hou. DreamStyle3D: Efficient 3D stylized asset generation via dual-attention disentanglement. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2026.

Peng Wang, Xiang Liu, and Peidong Liu. Styl3R: Instant 3D stylized reconstruction for arbitrary scenes and styles. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Fangzheng Wu and Brian Summa. Attention sinks in diffusion transformers: A causal analysis. In *International Conference on Machine Learning*, 2026.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3D: Scalable image-to-3D generation via 3D latent diffusion transformer. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121859–121881, 2024.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, and Yao Yao. Direct3D-S2: Gigascale 3D generation made easy with spatial sparse attention. In *Advances in Neural Information Processing Systems*, volume 38, pp. 170778–170804, 2025a.

Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3R: Amodal 3D reconstruction from occluded 2D images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9181–9193, 2025b.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14419–14429, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, pp. 21875–21895, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Bohan Zeng, Shanglin Li, Yutang Feng, Ling Yang, Juan Zhang, Hong Li, Jiaming Liu, Conghui He, Wentao Zhang, Jianzhuang Liu, Baochang Zhang, and Shuicheng Yan. IPDreamer: Appearance-controllable 3D object generation with complex image prompts. In *International Conference on Learning Representations*, 2025.

Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. ARF: Artistic radiance fields. In *European Conference on Computer Vision*, pp. 717–733, 2022.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. CLAY: A controllable large-scale generative model for creating high-quality 3D assets. *ACM Transactions on Graphics*, 43(4):1–20, 2024.

Jinghao Zhou, Tomas Jakab, Philip Torr, and Christian Rupprecht. Scene-conditional 3D object stylization and composition. In *European Conference on Computer Vision (ECCV)*, pp. 289–305, 2024.

Lin Zhu, Xinbing Wang, Chenghu Zhou, Qinying Gu, and Nanyang Ye. Less is more: Masking elements in image condition features avoids content leakages in style transfer diffusion models. In *International Conference on Learning Representations*, pp. 66094–66126, 2025.

Shihao Zhu, Ziheng Ouyang, Yijia Kang, Qilong Wang, Mi Zhou, Bo Li, Ming-Ming Cheng, and Qibin Hou. Mixture of style experts for diverse image stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30500–30510, 2026.

A IMPLEMENTATION DETAILS

Blank-image tokens. For TRELLIS.2, we encode an all-black 512×512 image through the pipeline's DINOv3 ViT-L/16 encoder (Siméoni et al., 2026) and parameter-free layer normalization. After discarding the CLS token and four registers, we store the 1,024 patch tokens, each of dimension 1,024, as a 4 MB tensor. Appending this cached tensor extends the reference sequence from 1,029 to 2,053 tokens. For the other backbones in Section 4.4, we match the blank image to the native background: black for TRELLIS, encoded with DINOv2 with registers (Darcet et al., 2024), and white for Hunyuan3D 2.1, encoded with DINOv2 (Oquab et al., 2024).

Stylization methods. Attention blending follows MorphAny3D's cross-attention rule (Sun et al., 2026). In each cross-attention block of the shape and texture SLat stages, latent queries Q attend separately to the content and reference, and the outputs are combined with style weight α:

$$\mathrm{Blend}_\alpha(Q) = (1-\alpha)\,\mathrm{Attn}(Q, K_c, V_c) + \alpha\,\mathrm{Attn}(Q, \widetilde{K}_r, \widetilde{V}_r), \tag{1}$$

where Attn is defined in Equation 2. The keys and values (K_c, V_c) come from the content tokens C_c, and $(\widetilde{K}_r, \widetilde{V}_r)$ from the augmented reference $\widetilde{C}_r = [C_r; B]$, or C_r without BACKDROP. MorphAny3D's implementation specifies the content weight, $1-\alpha$. Blending affects only the conditional pass of classifier-free guidance (Ho & Salimans, 2021); the sparse-structure stage uses the content alone. Table 3 summarizes the settings.

Our TRELLIS.2 implementations follow the released TRELLIS code (Qu et al., 2025; Sun et al., 2026). StyleSculptor's output stream attends to the reference stream's keys and takes its values only in the style-aware channels. We use intensity 4, the setting nearest $\alpha = 0.85$, which makes 80% of SLat channels and 7% of sparse-structure channels style-aware, with one-step unconditional refinement tails.

MorphAny3D adds temporally fused self-attention along a morphing chain, mixing each frame's self-attention output with attention to the previous frame's keys and values at weights 0.8 and 0.2. We use a 21-frame chain with $\alpha = k/20$ at frame k, sampling through frame 17 to reach $\alpha = 0.85$. With BACKDROP, all 17 sampled frames receive B because each depends on its predecessor. This shorter chain reduces the cost of the released 50-frame configuration, which requires sampling 42 frames per output. For all three methods, BACKDROP appends B only to the reference conditioning of the shape and texture SLat stages.

Stylize-then-generate pipelines. Both 2D stylizers receive a 1024×1024 content image composited onto white and the full-frame reference. IP-Adapter-Plus (Ye et al., 2023) uses SDXL (Podell et al., 2024) image-to-image with strength and adapter scale both set to 0.70, injecting the reference only into InstantStyle's style block (Wang et al., 2024). StyleExpert (Zhu et al., 2026) uses its released FLUX.1 Kontext configuration (Black Forest Labs et al., 2025), with `transformers` 4.56.2 and `diffusers` 0.35.1. The stylized image then passes through each backbone's standard image-to-3D pipeline, including Hunyuan3D-Paint 2.1 with its released settings for Hunyuan3D 2.1.

Sampling, hardware and timing. Each method retains its backbone's sampler and guidance settings. TRELLIS.2 uses guidance 7.5 on $t \in [0.6, 1]$ in the sparse-structure and shape stages, with no guidance in the texture stage. All runs use NVIDIA L40 GPUs. Times in Table 1 are means over the same 12 non-object pairs, one per content, from image preprocessing to the decoded asset, including Hunyuan3D 2.1 texturing. Stylize-then-generate times include both the 2D and 3D stages; model loading, warm-up, saving, rendering and scoring are excluded. A separate controlled comparison on 12 non-object and 12 object pairs isolates the cost of BACKDROP with attention blending: the longer sequence adds an average of 0.85 s (13%) to shape and texture sampling, with similar costs for both reference types, while decoding the larger intact meshes adds under 0.1 s.

B EVALUATION PROTOCOL DETAILS

Data. The 12 content images depict isolated characters, creatures, sculptures and everyday objects. Of the 36 non-object references, 20 come from 16 DTD categories (Cimpoi et al., 2014) and 8 from StyleBench (Gao et al., 2026); the remaining 8 are AI-generated or AI-edited images. They were chosen for recognizable geometric features, such as facets, folds, scales or lattices, and all methods use them full-frame without background removal. Each pair shares one seed across methods,

Method	Backbone (steps)	Weight setting	Stylized stages	BACKDROP tokens
Attention blending	TRELLIS.2 (12)	$\alpha = 0.85$	shape, texture	1,024 black
StyleSculptor	TRELLIS.2 (12)	intensity 4	all three	1,024 black
MorphAny3D	TRELLIS.2 (12)	frame 17 of 21	shape, texture	1,024 black
StyleSculptor	TRELLIS (25)	intensity 4	SS, SLat	–
MorphAny3D	TRELLIS (25)	frame 17 of 21	SLat	–
Attention blending	TRELLIS (12)	$\alpha = 0.85$	SLat	1,369 black
Attention blending	Hunyuan3D 2.1 (50)	$\alpha = 0.85$	shape	1,369 white
IP-Adapter-Plus	SDXL (50)	0.85	content image	–
StyleExpert	FLUX.1 Kontext (28)	none	content image	–

Table 3: Evaluation settings. Top: methods that stylize during generation (Table 1). Middle: cross-backbone attention-blending tests (Section 4.4). Bottom: 2D stylizers used before TRELLIS.2 or Hunyuan3D 2.1; IP-Adapter-Plus lists image-to-image strength and adapter scale. SS denotes sparse structure; TRELLIS has a single SLat stage. TRELLIS uses guidance 5.0 on $t \in [0.5, 1]$, and Hunyuan3D 2.1 uses 5.0 throughout.

including the 2D stylizers and unstylized base: the first 32 bits of the SHA-256 hash of its two file names.

For the object-reference analyses, we pair the same contents with 12 Objaverse-XL objects (Deitke et al., 2023a) rendered as cutouts. After TRELLIS.2 preprocessing, background occupies 277–664 of their 1,024 patches. These 144 pairs support the removal test (Figure 4 and Table 4) and the object-reference attention panels. The main evaluation uses the 432 non-object pairs; separate exploratory samples are described in Appendix F.

Rendering. Each output is rendered at 512×512 from three views, spaced 120° apart in yaw from the content's front at 20° elevation. TRELLIS.2 uses physically based rendering under a high-dynamic-range environment map; TRELLIS uses unlit vertex colors. Hunyuan3D 2.1 renders its textured assets, but geometry metrics use the saved shape-stage meshes because texturing includes remeshing. Its attention-blending tests in Section 4.4 stop at the shape stage, so we report no CSD for them.

Metrics. CSD averages the cosine similarity between the reference and the three output views using the released ViT-L style head (Somepalli et al., 2024). Inputs are RGB images, with renders on black, resized to a 224-pixel shorter side and center-cropped.

SIS measures three defects relative to the matched base: debris d, the increase in surface area fraction outside the largest assembly; fragmentation $f = \max(0, 1 - N_{\text{base}}/N)$; and open boundaries b, the increase in the fraction of edges bordering only one face. An assembly groups connected components within about 1% of the main component's diagonal of one another. The effective component count is $N = 1/\sum_i a_i^2$, where a_i is each component's area fraction, and N_{base} is the corresponding count for the base. We combine the three penalties as $\text{SIS} = \big((1-d)(1-f)(1-b)\big)^{1/3} \in [0, 1]$.

F@2% measures bidirectional point-to-mesh proximity. We normalize both meshes by the base's bounding box without alignment and sample 80,000 area-weighted points on each. Using a distance threshold of 2% of the base's longest dimension, precision measures the fraction of output points near the base mesh, and recall measures the reverse direction. F@2% is their harmonic mean, averaged over two sampling seeds. FID (Heusel et al., 2017), computed with clean-fid (Parmar et al., 2022), compares each method's 1,296 output views with its 432 reference images (36 distinct).

These metrics capture different aspects of the output. CSD combines shape and appearance, so it can reward reference-like features on broken geometry, such as holes resembling a honeycomb. SIS judges the surviving geometry, while F@2% detects missing parts but also penalizes intended shape changes. We therefore report them separately.

Aggregation. Table entries are means over all pairs; none produced an empty mesh, which would score zero on SIS and F@2%. For paired comparisons, we match outputs by content, reference and seed, average the differences across the 12 contents for each reference, and bootstrap the 36 reference means with 10,000 resamples. The 95% interval runs from the 2.5th to the 97.5th percentile.

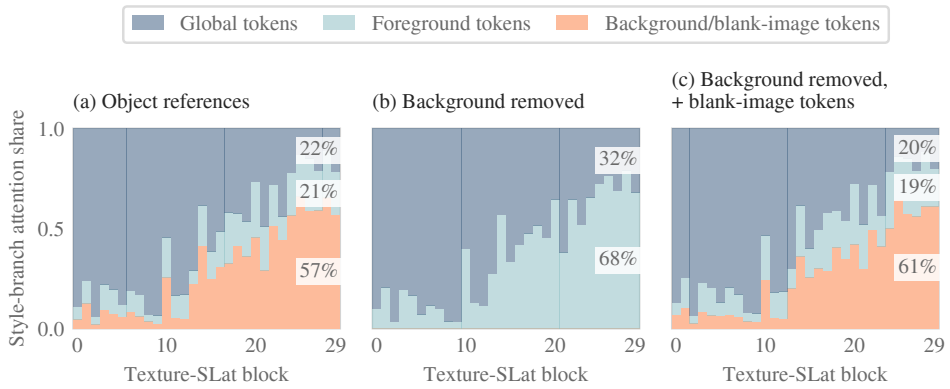


Figure 9: Reference-token attention across texture-SLat blocks with attention blending at $\alpha = 0.85$: (a) unmodified object references, (b) background tokens removed after encoding, and (c) background tokens replaced by blank-image tokens. Background or replacement tokens receive most deep-block attention. Each panel averages 144 object pairs, with token groups as in Figure 3.

Resampling at the reference level accounts for dependence among outputs that share a reference. Object-reference analyses use the same procedure with 12 references.

C Attention analysis and derivation

C.1 Measuring attention shares

We measure attention shares on the reference branch of attention blending, which operates only in the conditional pass of classifier-free guidance (Ho & Salimans, 2021). At each block, sampling step and head, we compute the model's attention weights after its query and key normalization and sum them within each token group. We then average over queries, heads and steps within each pair, and finally over pairs. This measurement leaves the model's projections unchanged. Each condition follows its own generation trajectory, so differences reflect both the conditioning tokens and the queries produced during generation.

For TRELLIS.2, we classify each 16×16 patch after cropping by the object's alpha mask and resizing to 512×512. A patch is background if it lies entirely outside the mask, including any padding, and foreground otherwise. Full-frame references have only foreground patches. The CLS token and four registers form the global group (Darcet et al., 2024). The other backbones use the same rule on 14×14 patches in a 518×518 image. TRELLIS has five global tokens and Hunyuan3D 2.1 only the CLS token; their shares average 12 and 50 sampling steps, respectively.

C.2 Texture stage

The texture stage shows the same shift from global tokens in early blocks to background tokens in deep blocks as the shape stage, though its final-block background share is lower (Figure 9). Removing background tokens redirects most of their attention to the foreground and the rest to global tokens. Replacing them with blank-image tokens restores a similar allocation. We did not measure the sparse-structure stage, which attention blending conditions on the content alone.

C.3 Blank-image tokens in cross-attention

Appending blank-image tokens redistributes attention between the original and appended groups. To derive this effect, let Q, K and V denote queries, keys and values, with head dimension d_h. Cross-attention is

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right) V, \tag{2}$$

where softmax normalizes each query's weights over the conditioning tokens. In both TRELLIS.2 SLat stages, Q and K include the model's root-mean-square normalization, applied separately to each token and head. Appending tokens therefore leaves the original keys unchanged.

For a fixed query q, each key k_i has logit $\ell_i = q^\top k_i/\sqrt{d_h}$ and value v_i. Let $\mathcal{R}$ and $\mathcal{B}$ index the original and appended tokens, with $Z_r = \sum_{i\in\mathcal{R}} e^{\ell_i}$ and $Z_b = \sum_{i\in\mathcal{B}} e^{\ell_i}$. Their combined attention output is

$$\begin{aligned}\mathrm{Attn}(q,[K_r;K_b],[V_r;V_b]) &= \frac{\sum_{i\in\mathcal{R}} e^{\ell_i} v_i + \sum_{i\in\mathcal{B}} e^{\ell_i} v_i}{Z_r + Z_b} \\ &= (1-m)\,\mathrm{Attn}(q,K_r,V_r) + m\,\mathrm{Attn}(q,K_b,V_b), \qquad m = \frac{Z_b}{Z_r+Z_b},\end{aligned} \tag{3}$$

where (K_r, V_r) and (K_b, V_b) are the original and appended keys and values. The appended group receives share m, scaling every original token's weight by $1-m$ while preserving their relative weights. This share varies with the query, head, block and step; Figures 3c and 9c show its averages. Setting $V_b = 0$, as in the zero-value ablation, leaves only $(1-m)\,\mathrm{Attn}(q, K_r, V_r)$. These identities describe a fixed query, not the changing queries along a generation trajectory.

D COMPLETE ABLATIONS

All comparisons use attention blending on TRELLIS.2 at $\alpha = 0.85$ unless stated otherwise; *clear* is defined in Section 4.1.

Removal and replacement on object references. Table 4 extends the test in Section 3.1, with CSD measured against the object references. Removing background tokens collapses nearly all outputs and lowers CSD. Replacing them with blank-image tokens yields no clear difference from the unmodified reference on any metric. Changing the background pixels before encoding also preserves structure for every tested fill, with similar final-block background attention (72–73%, versus 74% unmodified). Only white and random-color fills clearly lower CSD, alongside a shift in appearance. Appending blank-image tokens to the unmodified reference produces no clear change in CSD or SIS; this does not establish equivalence.

The removal test has two limits. Because removal follows encoding, the background has already influenced the remaining tokens through encoder self-attention. Token counts also differ: removal drops 277–664 tokens, while replacement adds 1,024.

Token controls. Replacing the blank-image tokens with copies of their mean removes all token variation and collapses 294 of 432 outputs, with no clear change in CSD relative to BACKDROP (Table 5). Their final shape-block attention falls from 71.0% to 41.7%, a paired difference of -29.3 percentage points (95% interval $[-30.1, -28.6]$). Because the conditions follow different generation trajectories, this association does not isolate attention magnitude as the cause of collapse. Keeping the blank-image keys but zeroing their values gives no clear SIS change and a small, clear CSD increase.

A fixed attention-share table. Building on the zero-value result, we test how much of the benefit a fixed reduction in reference attention can recover. We replace the query-dependent share m in Equation 3 with its mean under BACKDROP for each stage, block, sampling step and head. Scaling the reference output by $1-m$ then requires no appended tokens. This table recovers 90% of BACKDROP's SIS gain over attention blending and all of its mean CSD gain, but leaves 28 collapsed outputs versus 2 with BACKDROP (Table 6). SIS and F@2% remain clearly below both BACKDROP and the zero-value variant. A fixed reduction thus captures much of the benefit, although this test does not isolate the cause of the remaining gap. We tested only this table, estimated from the same 432 pairs on which it is evaluated.

Image borders. Table 7 compares each black border with its corresponding resizing control. Both borders preserve structure that is lost under resizing alone, supporting a role for the added background. Both also yield clearly lower CSD than BACKDROP at $\alpha = 0.85$; CSD was not measured at other weights. Separate measurements at $\alpha = 0.70$ show higher mean attention to both borders than to blank-image tokens at blocks 20 and 29. Encoding the border together with the subject also

Reference conditioning	CSD ↑	SIS ↑	F@2% ↑	Collapses
Unmodified (attention blending)	0.6181	0.9938	0.9905	0/144
Background removed	0.4070	0.0901	0.9358	143/144
Removed + blank-image tokens	0.6212	0.9931	0.9904	0/144
Background refilled before encoding				
White	0.5765	0.9949	0.9909	0/144
Subject's mean color	0.6161	0.9946	0.9908	0/144
Pixel noise	0.6018	0.9916	0.9904	0/144
Random-color patches	0.5997	0.9955	0.9909	0/144
Unmodified + BACKDROP	0.6139	0.9940	0.9913	0/144

Minus unmodified	ΔCSD	ΔSIS	ΔF@2%
Background removed	−0.2111	−0.9037	−0.0546
	[−0.2611, −0.1667]	[−0.9306, −0.8704]	[−0.0683, −0.0417]
Removed + blank-image tokens	+0.0031	−0.0007	−0.0000
	[−0.0070, +0.0134]	[−0.0037, +0.0024]	[−0.0003, +0.0002]
Background refilled before encoding			
White	−0.0416	+0.0011	+0.0005
	[−0.0594, −0.0241]	[−0.0004, +0.0029]	[−0.0001, +0.0014]
Subject's mean color	−0.0020	+0.0008	+0.0004
	[−0.0142, +0.0139]	[−0.0015, +0.0034]	[−0.0001, +0.0011]
Pixel noise	−0.0162	−0.0021	−0.0001
	[−0.0358, +0.0031]	[−0.0059, +0.0017]	[−0.0006, +0.0006]
Random-color patches	−0.0183	+0.0017	+0.0004
	[−0.0285, −0.0090]	[−0.0010, +0.0046]	[+0.0000, +0.0011]
Unmodified + BACKDROP	−0.0042	+0.0003	+0.0008
	[−0.0169, +0.0075]	[−0.0025, +0.0032]	[+0.0001, +0.0017]

Table 4: Background removal, replacement and refilling on object references, with attention blending on TRELLIS.2 at $\alpha = 0.85$; 144 pairs (12 contents $\times$ 12 object references) per condition. Removal drops each reference's background tokens (277–664 of its 1,024 patch tokens) after encoding; replacement then appends the 1,024 blank-image tokens; each fill replaces the background pixels before encoding; the last row appends the blank-image tokens to the unmodified reference. CSD is measured against the object references (the unstylized base scores 0.341). Collapses count outputs with SIS < 0.5. Paired differences are relative to the unmodified condition; the second line of each gives its 95% interval from bootstrapping the 12 references.

changes its tokens relative to an all-black canvas, while the reference tokens on the larger canvas remain close to their encoding without the border.

E COMPLETE QUANTITATIVE RESULTS

The following tables expand on Table 1 and Figure 7, using the paired bootstrap procedure in Appendix B.

Table 8 reports the paired effects of BACKDROP and additional geometry measures. SIS and F@2% improve clearly for all three methods; CSD improves clearly for attention blending and StyleSculptor, with no clear change for MorphAny3D. All three methods with BACKDROP also score clearly higher on CSD than every stylize-then-generate pipeline.

Precision and recall distinguish two kinds of departure from the unstylized base. Without BACKDROP, high precision and lower recall indicate that surviving geometry stays near the base's surface while parts are lost. Stylize-then-generate pipelines reduce both, consistent with a change in the object's shape. Face counts provide a further check: outputs without BACKDROP average fewer than a tenth of the base's faces, while those with it remain much closer to the base's count.

Reference conditioning	CSD ↑	SIS ↑	F@2% ↑	Collapses
No appended tokens	0.2542	0.1283	0.9209	418/432
BACKDROP	0.3314	0.9746	0.9925	2/432
Copies of the mean token	0.3110	0.3907	0.9732	294/432
Blank-image keys, zero values	0.3415	0.9844	0.9915	1/432

Δ [95% CI]	Mean copies − BACKDROP	Zero values − BACKDROP
CSD	-0.0204 [-0.0432, $+0.0019$]	$+0.0101$ [$+0.0027$, $+0.0176$]
SIS	-0.5839 [-0.6406, -0.5257]	$+0.0099$ [-0.0010, $+0.0247$]
F@2%	-0.0193 [-0.0244, -0.0147]	-0.0010 [-0.0013, -0.0006]

Table 5: Token-construction controls with attention blending on TRELLIS.2, 432 non-object pairs at $\alpha = 0.85$. Every augmented condition appends 1,024 tokens to both shape and texture stages. Mean copies remove token variation; zero values retain the blank-image keys while zeroing their projected attention values. Collapses count outputs with SIS < 0.5. Paired differences are relative to BACKDROP, with 95% intervals from bootstrapping the 36 references. These are fixed-strength comparisons; mean copies also remove the spread of the keys and values.

Reference conditioning	CSD ↑	SIS ↑	F@2% ↑	Collapses
No appended tokens	0.2542	0.1283	0.9209	418/432
BACKDROP	0.3314	0.9746	0.9925	2/432
Blank-image keys, zero values	0.3415	0.9844	0.9915	1/432
Fixed attention-share table	0.3328	0.8901	0.9890	28/432

Δ [95% CI]	Table − BACKDROP	Table − zero values
CSD	$+0.0013$ [-0.0070, $+0.0102$]	-0.0088 [-0.0177, $+0.0002$]
SIS	-0.0845 [-0.1189, -0.0535]	-0.0943 [-0.1272, -0.0652]
F@2%	-0.0035 [-0.0043, -0.0028]	-0.0025 [-0.0033, -0.0019]

Table 6: A fixed attention-share table in place of appended tokens, with attention blending on TRELLIS.2, 432 non-object pairs at $\alpha = 0.85$. The table appends nothing: in every head it scales the reference branch's attention output by $1 - m$, where m is the blank-image tokens' mean share of that branch's attention recorded for the same stage, block, sampling step and head in the BACKDROP runs on these pairs. Collapses count outputs with SIS < 0.5. Paired differences use 95% intervals from bootstrapping the 36 references.

The geometry gains hold across reference categories: SIS and F@2% improve clearly for every method in each category (Table 9). CSD gains are less uniform. All three methods improve clearly on scenes, and attention blending and StyleSculptor also improve on textures; none shows a clear change on the eight artistic images.

The sweep shows how this benefit depends on style weight (Table 10). At $\alpha = 0.40$ and 0.55, BACKDROP clearly lowers CSD, consistent with blank-image tokens taking a share of reference attention. The difference is unresolved at 0.70 and clearly favors BACKDROP at 0.85 and 1.00. At the default weight of 0.85, BACKDROP also exceeds the baseline's peak CSD, reached at 0.70, by $+0.062$ [$+0.038$, $+0.087$].

F ADDITIONAL QUALITATIVE RESULTS

To show appearance and geometry together, we split output views diagonally between a textured render (upper left) and surface normals (lower right), except where noted.

Gallery. Across all three methods, BACKDROP preserves recognizable parts where the unaugmented outputs fragment or become perforated shells (Figure 10). The references affect geometry as well as color and material: surface normals reveal leafy lobes on the chicken, hexagonal cells on the pineapple, a perforated shell on the robot crab and bumps on the bottle. MorphAny3D with BACKDROP stays closest to the unstylized base, consistent with its highest F@2% (Table 1). Preserving the

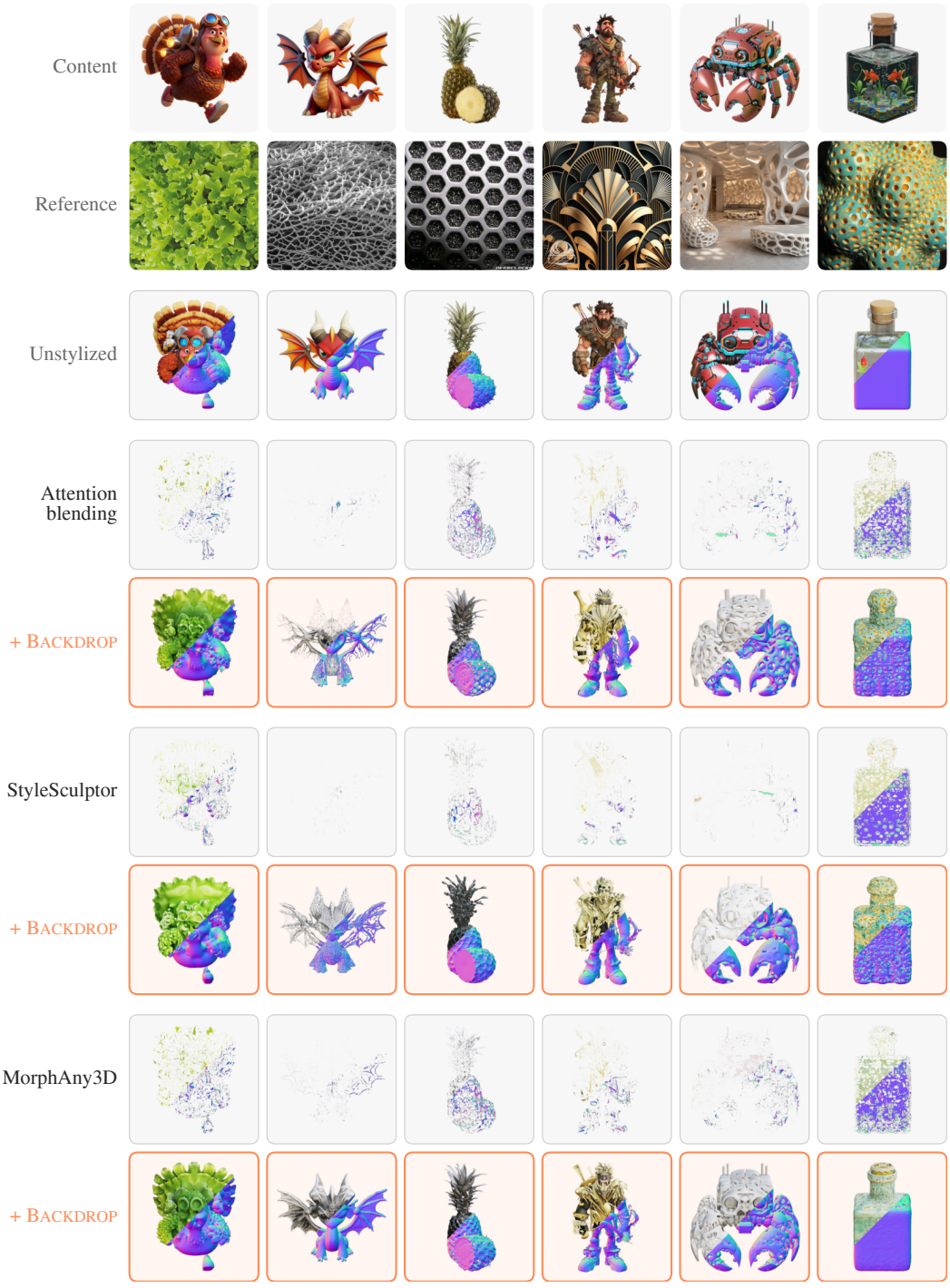


Figure 10: Six additional pairs from the TRELLIS.2 evaluation at $\alpha = 0.85$, one pair per column, with a shared camera and crop for each pair's outputs. None of these pairs appears in the main text. The references are four DTD textures, one StyleBench image and one scene.

Reference conditioning	CSD ↑	SIS ↑	F@2% ↑	Collapses
BACKDROP	0.3314	0.9746	0.9925	2/432
Border on a larger canvas	0.3141	0.9799	0.9926	2/432
Larger canvas, no border	0.2486	0.1384	0.9165	414/432
Border around a shrunken reference	0.3020	0.9903	0.9931	0/432
Shrunken reference, no border	0.2537	0.1250	0.9219	420/432

Paired contrast	ΔCSD	ΔSIS	ΔF@2%
Minus BACKDROP			
Border on a larger canvas	−0.017	+0.005	+0.0002
	[−0.027, −0.007]	[−0.001, +0.013]	[−0.0002, +0.0006]
Larger canvas, no border	−0.083	−0.836	−0.0759
	[−0.116, −0.050]	[−0.866, −0.803]	[−0.0958, −0.0580]
Border around a shrunken reference	−0.029	+0.016	+0.0006
	[−0.043, −0.017]	[+0.005, +0.029]	[+0.0003, +0.0010]
Shrunken reference, no border	−0.078	−0.850	−0.0706
	[−0.110, −0.046]	[−0.875, −0.822]	[−0.0873, −0.0553]
Border minus the same resizing without a border			
Border on a larger canvas	+0.065	+0.842	+0.0761
	[+0.033, +0.098]	[+0.808, +0.872]	[+0.0580, +0.0960]
Border around a shrunken reference	+0.048	+0.865	+0.0713
	[+0.013, +0.084]	[+0.840, +0.888]	[+0.0559, +0.0880]

Table 7: Image borders and matched image-size controls with attention blending on TRELLIS.2, 432 non-object pairs per setting at $\alpha = 0.85$. The border on a larger canvas surrounds the full-resolution reference on a 704×704 canvas (912 border tokens); the border around a shrunken reference surrounds a 368×368 reference on a 512×512 canvas (495 border tokens). The two controls apply the same resizing without a black border: the larger canvas is filled by the enlarged reference and encoded as 1,936 tokens, and the shrunken reference is resized back to the full canvas. Collapses count outputs with SIS < 0.5. The second line of each paired contrast gives its 95% interval from bootstrapping the 36 references.

outer shape does not guarantee the original appearance: under all three methods, the bottle becomes opaque and hides the fish inside.

Blank-image tokens in the texture stage. Holding the generated shape and texture-stage noise fixed isolates the effect of blank-image tokens on texture (Figure 11). Without them, the goblin and the robot crab turn a brighter green, and the dragon's bark and the boots' sausage skin take on more saturated reds and oranges. These observations are qualitative; we did not measure exposure or saturation.

Hunyuan3D 2.1. Attention blending at $\alpha = 0.85$ generates geometry from the reference with little preservation of the content: the honeycomb texture becomes a cube and the dragon reference a dragon (Figure 12a). SIS remains high, but F@2% is below 0.5 for all 1,152 outputs across object and non-object pairs, with and without blank-image tokens. This failure prevents testing the predicted repair on this backbone (Section 4.4).

We also remove background tokens after encoding during direct generation from the 12 content images and 12 object references, using one seed each (Figure 12b). All content images still produce their objects, generally coarser and inflated. Eight object references remain intact but coarser, three lose structure and one collapses to its ground plate. Background tokens therefore affect geometry here too, though this exploratory single-image test addresses a different task from stylization.

F.1 DIRECT GENERATION FROM NON-OBJECT INPUTS

With attention blending, the content image supplies the coarse voxel support, while BACKDROP augments reference conditioning in the later stages. Direct generation from a texture, scene or painting instead requires TRELLIS.2 to infer the coarse geometry from that image alone. We explored this

Method	ΔCSD [95% CI]	ΔSIS [95% CI]	ΔF@2% [95% CI]
Attention blending	+0.077 [+0.044, +0.110]	+0.846 [+0.819, +0.871]	+0.072 [+0.055, +0.090]
StyleSculptor	+0.079 [+0.046, +0.112]	+0.799 [+0.769, +0.829]	+0.108 [+0.078, +0.143]
MorphAny3D	+0.023 [−0.006, +0.052]	+0.799 [+0.764, +0.832]	+0.025 [+0.019, +0.033]

Method	Backbone	Collapses	Faces (M)	Precision	Recall
Unstylized base	TRELLIS.2	–	3.97	–	–
Unstylized base	Hunyuan3D 2.1	–	0.99	–	–
Attention blending	TRELLIS.2	418/432	0.21	0.960	0.891
+ BACKDROP	TRELLIS.2	2/432	3.56	0.993	0.992
StyleSculptor	TRELLIS.2	406/432	0.19	0.943	0.826
+ BACKDROP	TRELLIS.2	7/432	5.33	0.977	0.978
MorphAny3D	TRELLIS.2	391/432	0.31	0.986	0.959
+ BACKDROP	TRELLIS.2	0/432	3.62	0.997	0.997
IP-Adapter-Plus → 3D	TRELLIS.2	0/432	2.88	0.608	0.577
StyleExpert → 3D	TRELLIS.2	0/432	4.13	0.642	0.667
IP-Adapter-Plus → 3D	Hunyuan3D 2.1	0/432	0.95	0.503	0.488
StyleExpert → 3D	Hunyuan3D 2.1	0/432	1.07	0.556	0.575

Table 8: Paired effect of BACKDROP and geometric checks behind Table 1: 432 non-object pairs per row, methods that stylize during generation at $\alpha = 0.85$ or its counterpart. Top: differences with minus without BACKDROP on TRELLIS.2, with 95% intervals from bootstrapping the 36 references. Bottom: collapses (outputs with SIS below 0.5), mean face count in millions, and the mean precision and recall that make up F@2%. Hunyuan3D 2.1 geometry is measured on its shape-stage meshes.

	Textures (23)			Scenes (5)			Artistic images (8)		
Method	CSD	SIS	F@2%	CSD	SIS	F@2%	CSD	SIS	F@2%
Attention blending	0.257	0.104	0.9177	0.248	0.208	0.9259	0.248	0.148	0.9272
+ BACKDROP	0.333	0.984	0.9938	0.428	0.976	0.9867	0.266	0.946	0.9923
StyleSculptor	0.251	0.148	0.8504	0.210	0.245	0.9199	0.236	0.201	0.8902
+ BACKDROP	0.331	0.975	0.9756	0.394	0.993	0.9800	0.250	0.949	0.9804
MorphAny3D	0.272	0.159	0.9707	0.220	0.226	0.9711	0.230	0.255	0.9740
+ BACKDROP	0.286	0.990	0.9972	0.354	0.993	0.9959	0.208	0.986	0.9970

Table 9: Results by reference category on TRELLIS.2 at $\alpha = 0.85$, averaged over the 12 contents and each category's references (23 textures, 20 of them from DTD; 5 scenes; 8 artistic images from StyleBench). BACKDROP keeps nearly all outputs intact in every category; its CSD gain is largest on scenes and not clear on the artistic images.

setting on 53 non-object images with two seeds each, outside the main evaluation protocol; we report illustrative outcomes, not a success rate.

With blank-image tokens in the shape and texture stages, some inputs yield appropriate geometry: houses and trees from a wool-craft village at one seed, or a coiled relief from a spiral texture (Figure 13a). Others produce picture panels, textured boxes or fragments, with outcomes varying by seed. An intact surface need not represent the intended geometry, although a thin relief can be appropriate for a material photograph.

Adding blank-image tokens to the sparse-structure stage changes this coarse interpretation, sometimes for better and sometimes for worse. At the first seed, perforated metal becomes a coherent sheet, but the village becomes a picture panel. Later stages build on the chosen support and do not reliably correct it. Neither changes to sparse-structure guidance nor token timing produced appropriate geometry consistently across the explored inputs and seeds.

A single early update can change the outcome (Figure 13b). Two village generations share their initial noise and first two sparse-structure updates with blank-image tokens. Removing the tokens from the third update onward yields houses and trees; retaining them for one more update yields a

α	Method	CSD ↑	SIS ↑	F@2% ↑	Collapses	ΔCSD [95% CI]
0.40	Attention blending	0.171	0.928	0.9978	12/432	
	+ BACKDROP	0.153	0.997	0.9985	0/432	−0.019 [−0.026, −0.011]
0.55	Attention blending	0.243	0.645	0.9937	155/432	
	+ BACKDROP	0.207	0.994	0.9973	0/432	−0.036 [−0.053, −0.019]
0.70	Attention blending	0.269	0.270	0.9746	355/432	
	+ BACKDROP	0.279	0.990	0.9955	0/432	+0.010 [−0.016, +0.035]
0.85	Attention blending	0.254	0.128	0.9209	418/432	
	+ BACKDROP	0.331	0.975	0.9925	2/432	+0.077 [+0.044, +0.110]
1.00	Attention blending	0.233	0.120	0.8334	423/432	
	+ BACKDROP	0.353	0.935	0.9891	16/432	+0.120 [+0.083, +0.155]

Table 10: Style-weight sweep with attention blending on TRELLIS.2 (Figure 7), 432 non-object pairs per row. Collapses count outputs with SIS below 0.5. ΔCSD is BACKDROP minus attention blending alone at the same weight, with 95% intervals from bootstrapping the 36 references.

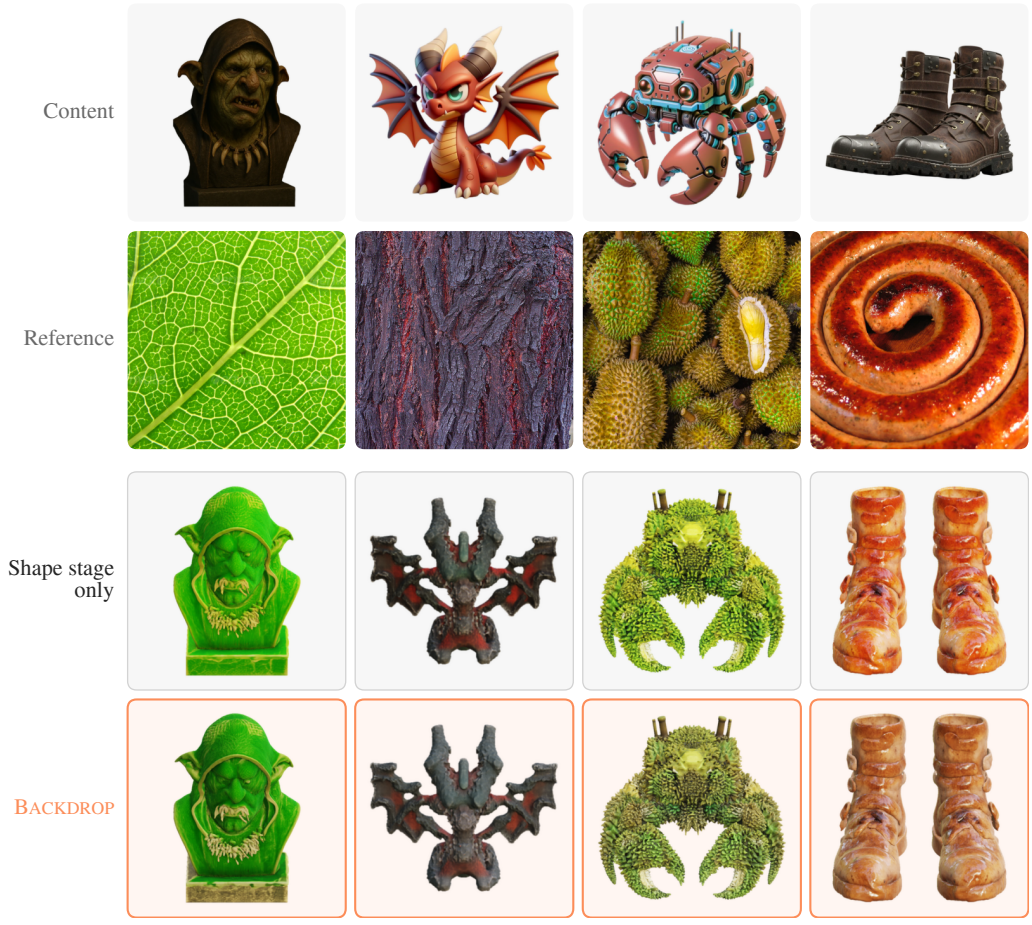
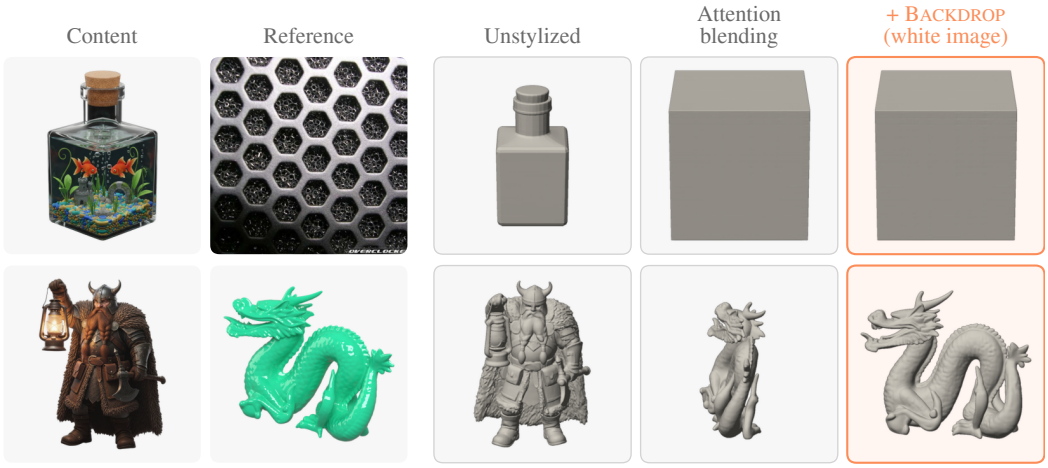


Figure 11: Texture-stage comparison with attention blending on TRELLIS.2 at $\alpha = 0.85$, shown as textured front views. Both conditions use the same shape and texture noise. Skipping the blank-image tokens in the texture stage results in brighter, oversaturated textures.

picture panel that persists through the remaining nine updates. In this example, the difficulty lies in choosing the coarse geometry, which blank-image tokens alone do not resolve.

(a) Attention blending on Hunyuan3D 2.1, $\alpha = 0.85$, shape stage only



(b) Direct generation on Hunyuan3D 2.1 without stylization, one seed per image

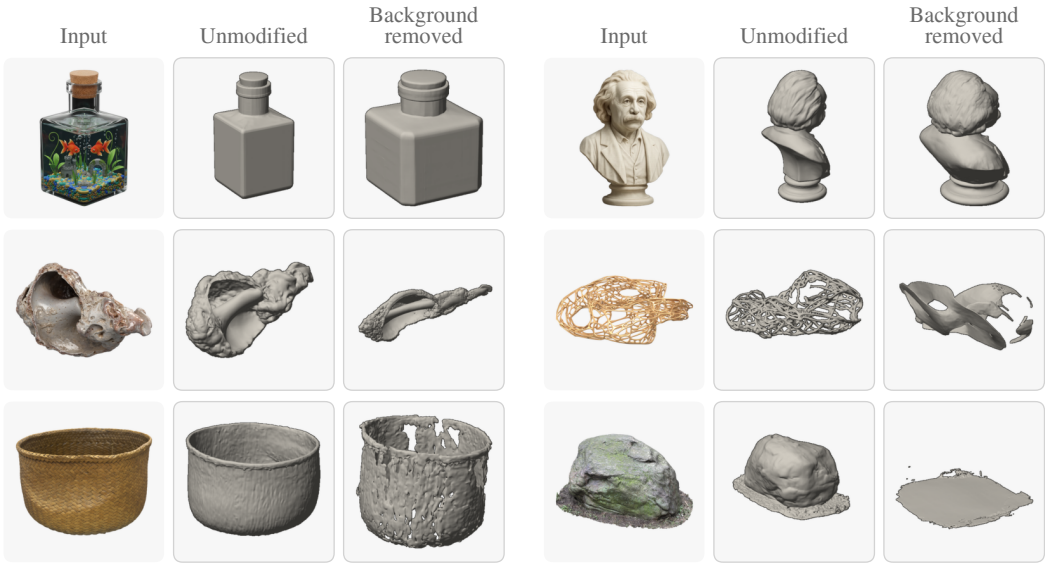
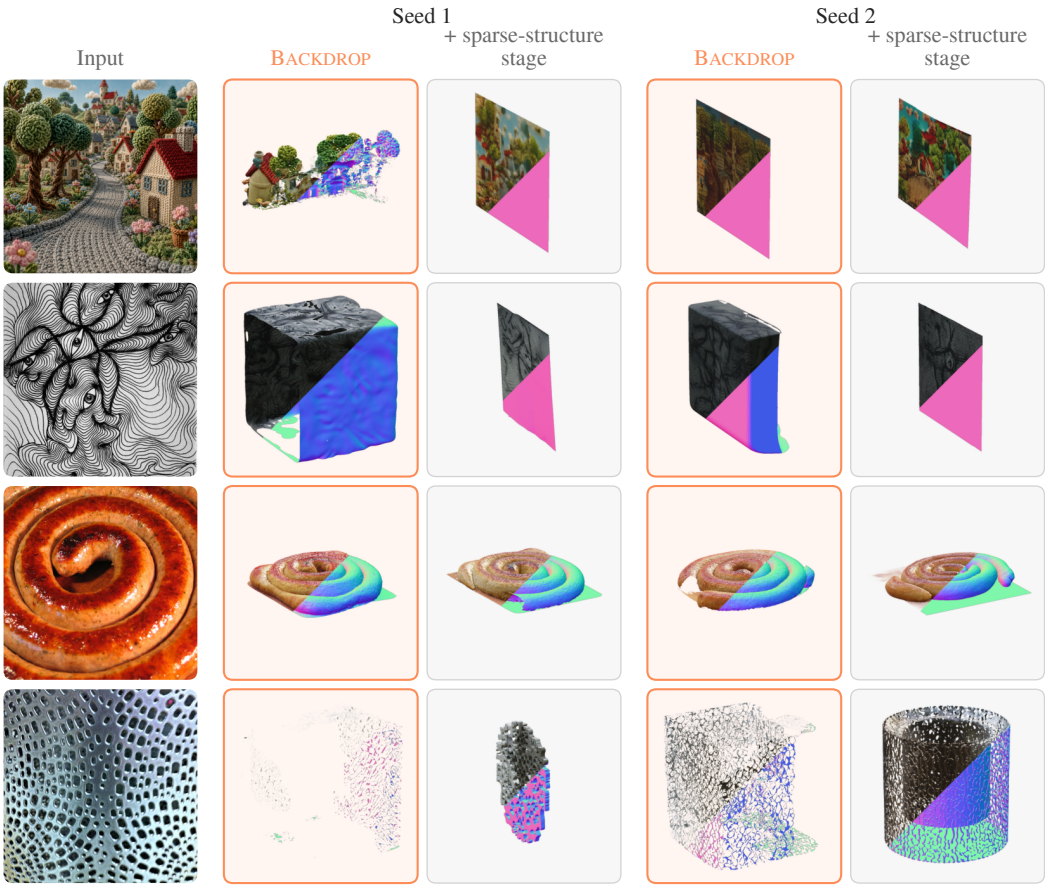


Figure 12: Hunyuan3D 2.1 shape-stage meshes, untextured, with one seed per example. (a) Two manually selected attention-blending examples at $\alpha = 0.85$, without and with tokens from an all-white image. (b) Direct generation before and after background-token removal: the two content images with the greatest relative volume increase and the four object references that lose structure or collapse.

(a) Direct generation on TRELLIS.2 from non-object images



(b) Wool-craft village, seed 1: blank-image tokens in the first k of 12 sparse-structure updates



Figure 13: Direct generation from non-object images on TRELLIS.2, shown in oblique views. (a) Blank-image tokens in the shape and texture stages only, or also in sparse structure, with two seeds. Rows illustrate the outcomes discussed in the text and do not indicate their frequency. (b) Village generations with shared initial noise and blank-image tokens in the first k of 12 sparse-structure updates; $k = 0$ and $k = 12$ repeat the first row's seed-1 outputs in (a).