DISENTANGLING TEXTURE AND COLOR FOR CONTROLLABLE 3D STYLIZATION AND MORPHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Separate texture and color references do not guarantee separate effects in a pretrained 3D generator: both attributes share a nonlinear decoder, and texture editing changes the state on which color acts. We formulate selective control as carrier-conditioned editing in TRELLIS structured latent space. Detail-oriented conditioning first constructs a textured carrier on the source's sparse support. Color is then fitted separately to the source and textured carriers, with preservation defined relative to each carrier. The four fitted endpoints define a two-attribute latent surface whose interaction term adapts the color direction to texture strength and exactly recovers all four corners. No pretrained weights are updated. We completed 120 balanced content–texture–palette triplets and exported 480 endpoint GLBs. A matched ten-triplet ablation measures the texture-transfer/content-preservation trade-off, and decoded control grids expose the effect of the interaction term. These results support endpoint generation and inspectable two-axis control, while matched external baselines and reference-based interior evaluations remain necessary to establish comparative superiority and independent decoded effects.

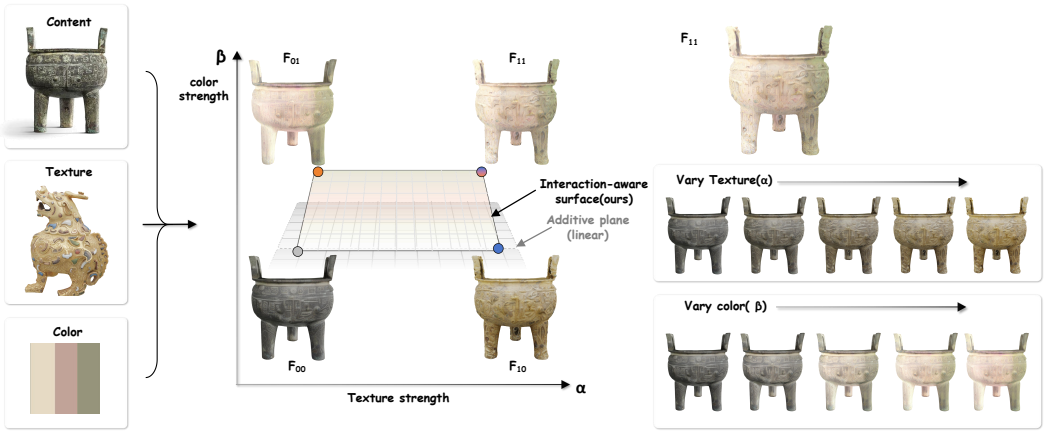


Figure 1: **Separate references, carrier-conditioned endpoints, and two-axis control.** A Chinese ding is edited into source F_{00}, texture-only F_{10}, color-only F_{01}, and joint F_{11} states. The strips show decoded texture and color sweeps. The central surface is conceptual, not a measured geometric plot.

1 INTRODUCTION

Reference-guided 3D stylization enables the same object to be adapted to different materials, colors, and visual themes. Appearance transfer has been studied for radiance fields, Gaussian representations, and meshes (Zhang et al., 2022; Liu et al., 2024; Xie et al., 2024); pretrained generators additionally support reference-guided manipulation of texture and geometry (Qu et al., 2025). Yet separating the reference inputs does not necessarily separate their effects in the generated representation. In the structured latent space of TRELLIS (Xiang et al., 2025), the same features jointly influence geometry and appearance, so a color edit may also alter local contrast, shading, or relief that contributes to perceived texture.

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.

Texture editing also changes the latent operating point, so a color displacement fitted on the original object may induce a different decoded effect after texture transfer. This makes the assumption of a fixed, transferable color offset overly restrictive for compositional editing. Continuous control further complicates the problem. Structured-latent morphing supports smooth transitions between 3D states (Sun et al., 2026), but a single interpolation coefficient does not provide independent texture and color control. As illustrated in Fig. 1, composing four attribute-specific endpoints therefore requires accounting for how the color response changes with the underlying carrier.

We address this problem through *carrier-conditioned* editing. We first construct aligned source and texture carriers on a shared sparse support, suppressing low-frequency color bias in the texture reference while retaining local luminance and chromatic variations. We then initialize a color displacement and refine it independently for each carrier, using the source carrier to preserve the original state during color-only editing and the texture carrier to preserve transferred texture during joint stylization. This formulation promotes selective editing without assuming that texture and color correspond to globally independent latent directions.

The four fitted states also expose a finite difference between the source-conditioned and texture-conditioned color edits. Retaining this interaction produces a bilinear latent surface with two control coordinates and exact endpoint recovery. Intermediate states are decoded from this surface without refitting, although decoding is still required and neither perceptual monotonicity nor fixed geometry follows from latent interpolation alone.

Contributions. (1) We formulate texture–color control as carrier-conditioned inverse editing, combining detail-oriented carrier construction with color fitting constrained by the current carrier's decoded geometry and appearance. (2) We retain the difference between the two fitted color displacements in a two-attribute latent surface, yielding a texture-dependent color direction and recovering all four endpoints.

Evidence and scope. We distinguish generation coverage (120 triplets, 480 endpoint assets) from quantitative image and geometry evaluation (ten triplets), matched implementation ablations (the same ten inputs), and a single-case interaction comparison. Main-text evidence emphasizes these controlled or explicitly scoped measurements. Cross-method numbers from a different evaluation set are reported as context in Table 1, not used to assert a matched-split ranking.

2 Related Work

Attribute-specific stylization. Feature-statistic transfer (Gatys et al., 2015; Huang & Belongie, 2017; Li et al., 2017) and attention-based conditioning (Park & Lee, 2019; Ye et al., 2023; Hertz et al., 2024) provide foundations for reference-guided appearance editing. Attribute-specific approaches distinguish color from style (Agarwal et al., 2024); SADis (Qin et al., 2025) uses color–grayscale differences for separate texture and color references. We inherit paired conditioning as an initialization, then fit a shared 3D latent under carrier-relative decoded constraints rather than treat the difference as an exact color factor.

3D appearance control. ARF, StyleRF, and StyleGaussian edit appearance in a common 3D representation (Zhang et al., 2022; Liu et al., 2023a; 2024). ARF-Plus controls perceptual factors, and StyleTex generates reference-guided texture on a given mesh (Li et al., 2025; Xie et al., 2024). StyleSculptor (Qu et al., 2025) is a close generative comparison for texture/geometry guidance; OrnaStyler (Aizawa et al., 2026) studies ornament-aware latent editing. Our distinction is carrier-relative color fitting followed by joint latent control, not the broader idea of training-free 3D stylization.

Our use of a shared generative state differs from editing an already fixed appearance representation. Direct color assignment or optimizing only Gaussian colors can preserve geometry by construction. Carrier-relative latent fitting instead asks whether a common editable state can retain the current carrier's details while attaining an external palette and supporting subsequent two-attribute morphing. This choice motivates a preservation–attainment comparison; a shared latent representation alone is not evidence of better appearance transfer.

Structured-latent morphing. TRELLIS (Xiang et al., 2025) supplies our frozen generator and decoder; MorphAny3D (Sun et al., 2026) transitions between objects by blending structured latents.

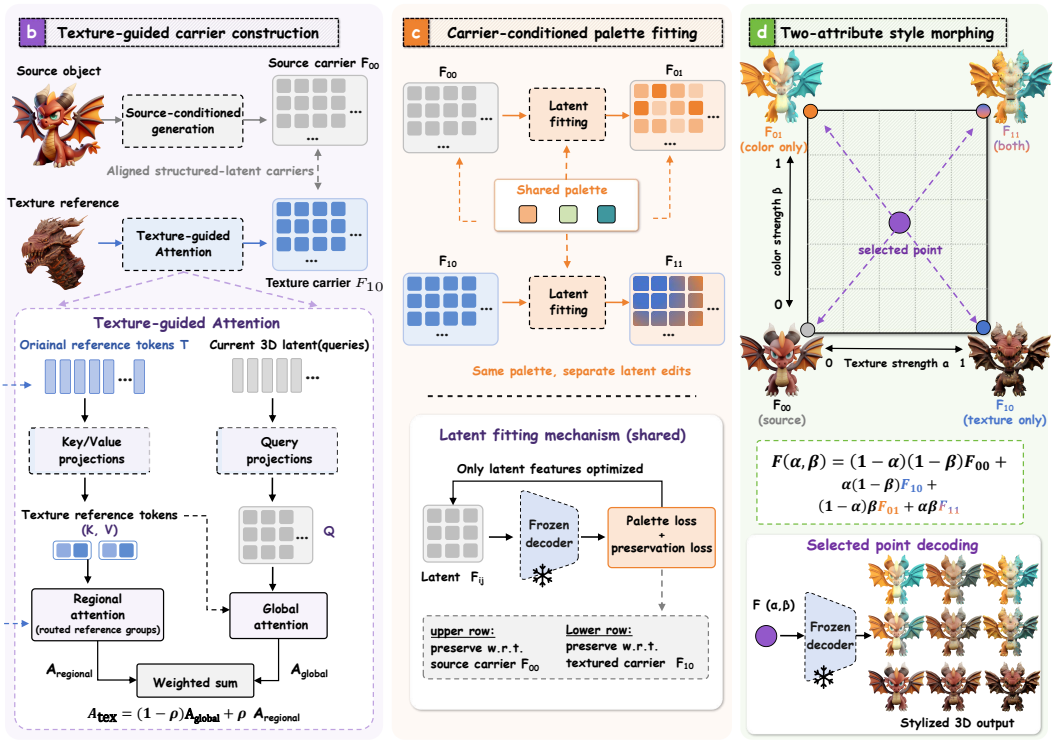


Figure 2: **Carrier construction, separate palette fitting, and latent composition.** The two fitting branches share a palette but preserve different carriers: F_{00} for F_{01} and F_{10} for F_{11}. Only latent features are optimized; decoder weights remain frozen. F_{10} is the *texture carrier*. Reference preparation (panel a of the full diagram) is enlarged in Appendix B.1.

Our domain instead spans texture and color strengths on one support and retains two carrier-specific color edits in a standard bilinear surface. Appendix A gives extended background.

3 Method

3.1 Problem and Shared Structured Latents

Given a source image I_s, texture reference I_t, and color reference I_c, we seek a family $\mathcal{O}(\alpha, \beta) = D_S(F(\alpha, \beta))$, where $\alpha, \beta \in [0, 1]$ control texture and color strength, respectively. TRELLIS first predicts occupied voxel coordinates $S = \{s_i\}_{i=1}^N$ and then samples features $F \in \mathbb{R}^{N \times d}$ on this support. We denote the image-conditioned sampler by $F = \mathcal{G}_\theta(S, E(I), \epsilon)$ and the frozen decoder by D_S. The four states F_{00}, F_{10}, F_{01}, and F_{11} share the same support S, preserving latent correspondence without enforcing identical decoded geometry.

We use *texture* to denote local patterns and relief, and *color* to denote chromaticity and low-frequency luminance. Since these attributes overlap perceptually, independent latent directions do not necessarily produce independent decoded effects. Locally,

$$D_S(F + \Delta F) \approx D_S(F) + J_D(F)\Delta F,$$

and the response depends on the current latent state F. Rather than explicitly computing the decoder Jacobian or its nullspace, we measure preservation relative to the carrier on which the color edit is applied (Fig. 2).

3.2 Detail-Oriented Texture Carrier Construction

Frequency-selective conditioning. We construct F_{10} on the source support using a reference that suppresses global color bias without discarding local chromatic detail. Let $\mathcal{P}_\sigma$ be Gaussian smoothing,

$\mathcal{H}_\sigma(I) = I - \mathcal{P}_\sigma(I)$ its high-pass residual, and Y the RGB-to-luminance mapping. We separate luminance and luminance-free chromatic variations:

$$H_Y = \mathcal{H}_\sigma(Y(I_t)), \qquad H_\chi = \mathcal{H}_\sigma(I_t) - Y(\mathcal{H}_\sigma(I_t))\mathbf{1}. \tag{1}$$

Here $\mathbf{1} \in \mathbb{R}^3$ broadcasts luminance across RGB channels. We reconstruct a neutral-luminance reference and then restore controlled chromatic detail:

$$\widetilde{I}_t^Y = \mathrm{clip}_{[r_-,r_+]}\big(\mu + \lambda_\ell \mathcal{N}_Y(H_Y)\big), \qquad \widetilde{I}_t = \mathrm{clip}_{[u_-,u_+]}\big(\widetilde{I}_t^Y \mathbf{1} + \lambda_\chi \mathcal{N}_\chi(H_\chi)\big). \tag{2}$$

The operators $\mathcal{N}_Y, \mathcal{N}_\chi$ normalize residuals; μ sets neutral luminance, $\lambda_\ell, \lambda_\chi$ control detail, and clipping bounds the output. Unlike grayscale conversion alone, this retains chromatic texture variation while attenuating low-frequency color bias. Appendix B.1 expands the preparation pathway.

Reference tokens and group routing. The frozen image encoder produces texture-reference tokens T. For grouping, we extract and ℓ_2-normalize only patch tokens, excluding prefix tokens. At the same patch locations, smoothed luminance, detail response, and normalized image coordinates are concatenated with these descriptors. Foreground patches are clustered into four appearance groups; background patches remain unassigned. These descriptors determine regional routing, not new key/value features. The original reference tokens supply (K_t, V_t), while queries Q come from the current sparse 3D latent:

$$A_{\text{global}} = \mathrm{softmax}(QK_t^\top/\sqrt{d_k})V_t, \qquad A_t = (1-\rho)A_{\text{global}} + \rho A_{\text{regional}}. \tag{3}$$

Global attention accesses the full reference; the regional response organizes access by the appearance groups, with mixing strength ρ. The groups are not guaranteed semantic parts, and image coordinates are grouping cues rather than explicit 2D–3D correspondences. Sampling on the same support S yields aligned source and texture carriers F_{00} and F_{10} without updating pretrained weights.

3.3 Carrier-Conditioned Color Fitting

Shared initialization, separate fits. Following paired color–grayscale conditioning (Qin et al., 2025), we sample auxiliary color, grayscale, and neutral latents $F_{\text{col}}, F_{\text{col},g}, F_{\text{col},n}$ with matched noise, together with a grayscale source latent $F_{s,g}$. We form

$$\Delta C_{\text{init}} = \mathcal{V}\big(F_{\text{col}} - F_{\text{col},g},\, F_{\text{col},g} - F_{\text{col},n}\big) - (F_{00} - F_{s,g}). \tag{4}$$

where $\mathcal{V}$ suppresses gray-to-neutral variation to reduce luminance leakage. The source pair is not exactly matched along one sampling trajectory, so this residual is not an exact color factor. Let

$$\Delta T = F_{10} - F_{00}, \qquad T_H = \Delta T - \mathcal{S}_{r_t}(\Delta T), \qquad \Delta C = \mathcal{S}_{r_c}(\Delta C_{\text{init}}).$$

where $\mathcal{S}_r$ smooths the sparse support. Texture-aware refinement suppresses the component along the observed high-frequency texture direction:

$$\widehat{\Delta C} = \Delta C - \mathrm{Proj}_{T_H}(\Delta C). \tag{5}$$

Projection is channel-wise and followed by normalization and RMS rescaling (Appendix B.2). Both branches start from the resulting residual,

$$F_{i1}^{(0)} = F_{i0} + \gamma\widehat{\Delta C}, \qquad C_i^* = \mathcal{T}\big(D_S(F_{i0}), I_c\big), \qquad i \in \{0,1\}. \tag{6}$$

where $\mathcal{T}$ transfers the target color distribution while adapting to each carrier's luminance and shading. The same I_c therefore specifies a common palette request, not identical decoder-space targets. For $i = 0$, the source determines the appearance to retain; for $i = 1$, the already textured carrier supplies that reference. Both C_i^* and the preservation anchor are constructed from F_{i0} before fitting and remain branch-specific. Residual refinement supplies a common starting direction; it does not constrain the two optimized displacements to remain equal.

Carrier-relative preservation. Starting from $F_{i1}^{(0)}$, we optimize only latent features with the frozen decoder:

$$\mathcal{L}_i(F) = \mathcal{L}_{\text{col}}\big(D_S(F), C_i^*\big) + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}}\big(D_S(F), D_S(F_{i0})\big) + \lambda_z d(F, F_{i0}), \qquad i \in \{0,1\}. \tag{7}$$

The first term seeks the carrier-specific color target; the second penalizes changes to the carrier's decoded appearance and geometry; the third limits latent deviation. The resulting fit gives F_{01} for

$i = 0$ and F_{11} for $i = 1$. In particular, F_{11} is optimized relative to F_{10}, not obtained by transferring the color displacement fitted around F_{00}. Preservation remains soft: a shared support and decoder penalties do not enforce fixed mesh geometry.

Why two independent fits? Transferring the source edit would set $F_{11} = F_{10} + (F_{01} - F_{00})$, reusing a displacement at a different decoder operating point. Instead, our joint branch preserves F_{10} rather than pulling the transferred texture toward the untextured F_{00}. Gradients pass through the frozen decoder only to the edited features; support, weights, and carrier anchors remain fixed. These are independent fits, not sequential recoloring steps. Additional endpoint examples are provided in Fig. F.1 of Appendix F.

3.4 Interaction-Aware Two-Attribute Morphing

Let $\Delta T = F_{10} - F_{00}$, $\Delta C_s = F_{01} - F_{00}$, and $\Delta C_t = F_{11} - F_{10}$. We define the interaction term as the mixed finite difference

$$K = \Delta C_t - \Delta C_s = (F_{11} - F_{10}) - (F_{01} - F_{00}), \tag{8}$$

which captures how the fitted color displacement changes after texture editing. We incorporate this interaction into the latent surface

$$F(\alpha, \beta) = F_{00} + \alpha \Delta T + \beta \Delta C_s + \alpha\beta K, \qquad (\alpha, \beta) \in [0, 1]^2. \tag{9}$$

When $K = 0$, the formulation reduces to additive composition with a fixed color direction. Equivalently, Eq. 9 can be written as

$$F(\alpha, \beta) = (1 - \alpha)(1 - \beta)F_{00} + \alpha(1 - \beta)F_{10} + (1 - \alpha)\beta F_{01} + \alpha\beta F_{11}. \tag{10}$$

Boundary recovery and carrier dependence. Equation (10) recovers all four fitted endpoints. Along $\beta = 0$, it interpolates from source to texture; along $\alpha = 0$, it interpolates from source to color. The opposite color boundary interpolates from F_{10} to its separately fitted F_{11}, rather than reusing the source color offset. More generally,

$$\frac{\partial F}{\partial \beta} = (1 - \alpha)\Delta C_s + \alpha \Delta C_t, \qquad \frac{\partial^2 F}{\partial \alpha \, \partial \beta} = K. \tag{11}$$

Thus K is the mixed finite difference of the anchors and the mixed derivative of the bilinear surface, not a learned interaction network. Relative to additive composition, the correction is exactly $\alpha\beta K$: it vanishes on the two source-adjacent boundaries and reaches K at the joint corner. This separates endpoint recovery, which follows algebraically, from the decoded interior behavior tested in Sec. 5.

Endpoint reuse. After fitting, each requested (α, β) combines cached endpoints through Eq. (9) and is decoded with D_S, without resampling or refitting. Shared support and latent continuity do not guarantee fixed decoded geometry, perceptual monotonicity, or independent attributes. Appendix C expands these identities and their interpretation.

4 Experiments

4.1 Experimental Setup

Datasets. We collect ten content images, twelve texture references, and fifteen color palettes from MorphAny3D (Sun et al., 2026) and StyleSculptor (Qu et al., 2025). Each content image is paired with every texture reference, while palettes are assigned cyclically for balanced usage, yielding 120 triplets. We generate all four endpoints for each triplet, resulting in 480 assets exported in Gaussian and GLB formats. Quantitative reference-agreement and geometry evaluations are conducted on a ten-triplet subset, with matched component ablations using the same inputs.

Implementation. We use the frozen TRELLIS image-large model with a fixed seed of 20260626. Sparse-structure, auxiliary SLAT, and texture-carrier sampling use 8, 8, and 80 steps, respectively. Color fitting is performed for 30 iterations with Adam (Kingma & Ba, 2015). For all ablations, the inputs and random seed are kept fixed. Appendix B provides the recorded configuration and additional implementation details.

Evaluation Metrics. Reference-agreement metrics compare the canonical-view 256×256 render of F_{11} with the corresponding input references. Content CLIP-I measures embedding cosine similarity (Radford et al., 2021). For color agreement, we report five-scale RGB MS-SWD and the Bhattacharyya distance between 8^3-bin HSV histograms (C-Hist), where lower is better for both. Texture agreement is evaluated against the grayscale texture reference using LPIPS (Zhang et al., 2018) and SSIM, with lower LPIPS and higher SSIM indicating closer agreement. Ablations additionally report texture CLIP-I and texture MS-SWD.

4.2 Experimental Results

Qualitative comparisons. Figure 3 qualitatively compares MorphAny3D, SADis+TRELLIS, StyleSculptor, and our four endpoints corresponding to source, color-only, texture-only, and joint edits. The selected examples illustrate appearance and shape changes under reference conditioning, rather than providing a matched quantitative ranking. MorphAny3D uses reference-image features for TRELLIS conditioning, which can couple geometry, color, and texture and lead to structural changes when transferring appearance. SADis+TRELLIS combines texture and color cues at the image level, but extending this formulation to sparse 3D representations can introduce additional coupling during reconstruction. StyleSculptor also exhibits mixed texture–color effects in some cases, with stylization tending toward one reference or blending cues from both. In contrast, our method first constructs a texture carrier and then performs carrier-conditioned color fitting, providing more explicit control over texture-only, color-only, and joint edits while reducing interference between the two attributes.

Table 1: **Quantitative results with existing 3D stylization methods.**

	Content	Color			Texture		
Method	CLIP-I↑	MS-SWD↓	C-Hist↓	VLM-C↑	LPIPS↓	SSIM$\times 10^3$ ↑	VLM-T↑
SADis + TRELLIS	0.73	18.38	0.99	37.17	0.65	116.10	18.00
StyleSculptor	0.74	21.26	0.89	35.44	0.69	124.61	23.22
MorphAny3D	0.71	18.31	0.91	40.05	0.65	117.21	19.72
Ours	0.778	17.502	0.848	88.64	0.649	81.84	31.29

Quantitative Results. Table 1 reports the quantitative comparison with prior methods. Our method achieves the highest content CLIP-I score of 0.778, indicating stronger semantic consistency with the content reference. For color fidelity, it obtains the lowest MS-SWD and C-Hist scores of 17.502 and 0.848, respectively, suggesting better agreement with the target color distribution. For texture consistency, our method achieves the lowest LPIPS score of 0.649, indicating higher perceptual similarity to the texture reference. Although its SSIM score is comparatively lower, the LPIPS result suggests that the transferred texture remains perceptually closer to the reference at the feature level. We further employ InternVL2.5-8B as a VLM-based evaluator for texture and color consistency. Our method achieves the highest VLM-C and VLM-T scores among the reported methods, further supporting its ability to preserve the specified color and texture attributes.

Geometry Preservation We compare each edited GLB against its corresponding F_{00} mesh on the same ten inputs. Each mesh is independently centered and normalized to a unit sphere, followed by uniform sampling of 100,000 surface points. Chamfer distance is computed as the sum of the two directed mean nearest-neighbor distances, while Hausdorff distance is the maximum of the two directed maxima. These normalized, sampled distances measure relative geometric deviation rather than ground-truth reconstruction error, and the independent normalization removes translation and scale differences.

As shown in Table 2, color-only fitting introduces the smallest geometric deviation, followed by texture-only and joint editing, consistent with texture affecting local surface structure. The self-resampling baseline reflects finite point sampling despite zero exact mesh self-distance and should therefore not be treated as a removable constant error. External comparisons are reported separately in Appendix D, since their samples and implementation settings are not directly matched.



Figure 3: **Qualitative comparison with prior methods.** Inputs are followed by MorphAny3D, SADis+TRELLIS, StyleSculptor, and our four fitted endpoints F_{00}, F_{01}, F_{10}, and F_{11}, corresponding to source, color-only, texture-only, and joint stylization.

Table 2: **Geometry preservation.** Chamfer and Hausdorff distances are reported in $\times 10^{-2}$.

Pair	Chamfer ↓	Hausdorff ↓
F_{00} vs. resampled F_{00}	1.03	2.04
Texture F_{10} vs. source F_{00}	1.38	6.56
Color F_{01} vs. source F_{00}	1.30	5.95
Joint F_{11} vs. source F_{00}	1.44	7.37

Average versus Extreme Deviations. Table 2 shows similar Chamfer distances across the edited endpoints (1.30–1.44), but larger variation in Hausdorff distance (5.95–7.37). Thus, comparable average geometric deviations can coexist with substantially different worst-case deviations, suggesting localized surface changes. Importantly, differences between source-relative distances, such as those of F_{10} and F_{11}, do not directly measure the geometric distance between the two edited meshes.

5 Ablations

Effect of Components. The same ten inputs and random seed are used for all variants in Table 3. Removing texture-reference attention decreases texture CLIP-I from 0.6505 to 0.6040, while content CLIP-I increases from 0.7783 to 0.8291 and color MS-SWD improves from 17.502 to 17.011, revealing a trade-off between texture transfer and source preservation rather than uniform degradation across all metrics. The remaining switches produce smaller, mixed effects. Removing content protection slightly worsens texture CLIP-I and LPIPS while improving C-Hist, whereas disabling channel-overlap handling slightly improves texture CLIP-I but worsens content CLIP-I, LPIPS, and both MS-SWD metrics.

Matched Ablation Analysis. Removing texture-reference attention decreases texture CLIP-I by 0.0465 while increasing content CLIP-I by 0.0508, indicating that its lower color MS-SWD comes with weaker texture transfer rather than uniformly better editing. The full variant achieves the lowest LPIPS, although the total LPIPS range is only 0.0016. Repeated runs with different seeds are therefore needed to determine whether these smaller differences are statistically reliable.

Interaction-Aware Interior Control. Figure 4 compares additive composition ($K = 0$, left) with the interaction-aware surface ($K \neq 0$, right) using the same content, texture, and color references.

Table 3: **Component ablation study.** All metrics evaluate F_{11} on the same ten inputs. The overlap variant sets the recorded channel-overlap coefficient to zero. Bold identifies a within-table optimum, not superiority over external methods.

Variant	Content CLIP-I ↑	Texture CLIP-I ↑	Texture LPIPS ↓	Color C-Hist ↓	Color MS-SWD ↓	Texture MS-SWD ↓
Full	0.7783	0.6505	**0.6490**	0.8479	17.502	30.419
No texture attention	**0.8291**	0.6040	0.6506	0.8645	**17.011**	**29.548**
No content protection	0.7788	0.6475	0.6496	**0.8465**	17.739	30.275
Overlap coeff. $= 0$	0.7732	**0.6508**	0.6504	0.8467	17.724	30.476

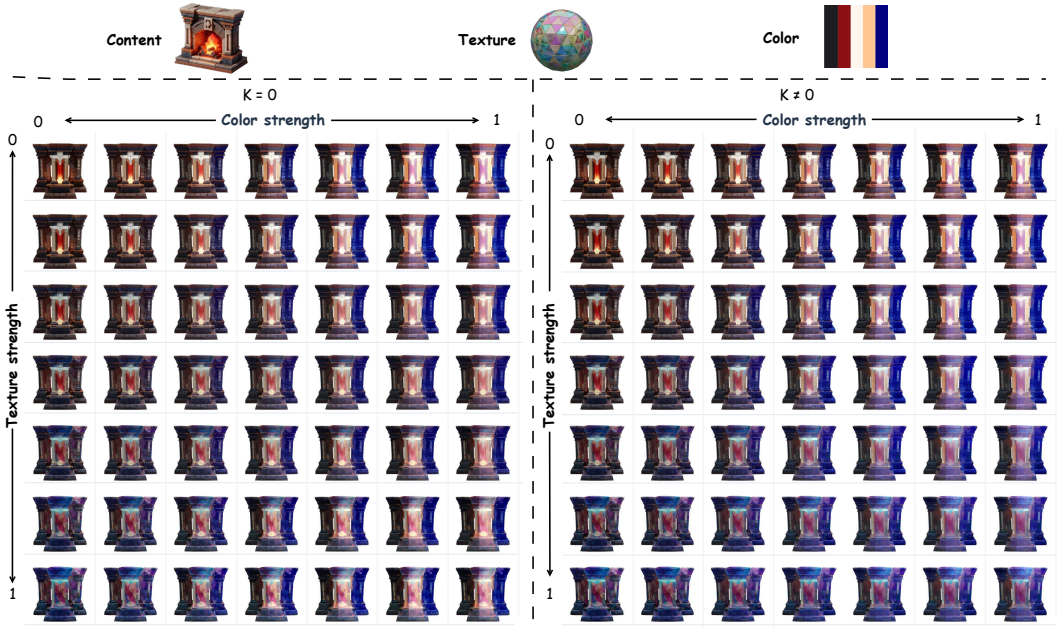


Figure 4: **Single-case interaction ablation.** Additive ($K = 0$, left) and interaction-aware ($K \neq 0$, right) 7×7 control grids share F_{00}, F_{10}, and F_{01}, while the joint corner differs by construction. Texture strength increases downward and color strength to the right.

Texture strength increases downward and color strength to the right in steps of $1/6$. Each cell is independently decoded from its latent state using the same sparse support, decoder, and camera, rather than image-space blending. Both surfaces share F_{00}, F_{10}, and F_{01}, whereas only the interaction-aware surface recovers F_{11} exactly. Table 4 reports adjacent-step diagnostics. Since setting $K = 0$ also changes the joint endpoint, this comparison characterizes the effect of the interaction term rather than isolating interior behavior under matched endpoints.

Interaction Strength Across the Grid. In the 7×7 grid, 13 boundary cells satisfy $\alpha\beta = 0$, where the interaction term vanishes and the two latent surfaces coincide. Across the 25 strict interior cells, the coefficient of K ranges from $1/36$ to $25/36$, so the RGB statistic measures a range of interaction strengths rather than repeatedly probing the joint endpoint. Because decoding is nonlinear, image-space differences need not scale linearly with these latent coefficients.

Table 4: **Decoded interaction diagnostics for the fireplace example.** RGB MAE is computed on the union foreground mask (any channel < 250) in the 0–255 range. Each reported step mean averages 42 adjacent pairs.

Variant	Texture-step MAE	Color-step MAE	Joint-corner deviation
Full surface ($K \neq 0$)	8.52	8.67	0.00
Additive ($K = 0$)	8.77	8.99	36.40

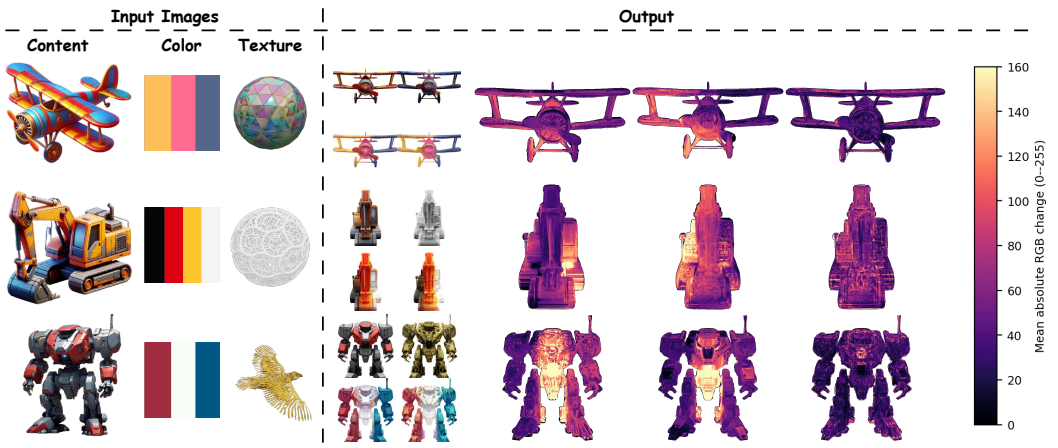


Figure 5: **Image-space color-response diagnostics.** Rows: biplane, excavator, and robot. Heatmaps show source-carrier and textured-carrier color responses and their difference, not latent K or a pure color factor. Geometry, shading, and alignment also affect these maps.

Interpreting the Diagnostics. Section 3.4 explains why the two surfaces coincide along boundaries where $\alpha\beta = 0$. Lower adjacent-step MAEs indicate smaller image changes, but do not by themselves imply smoother or more faithful control; reference-based evaluation is needed to distinguish these possibilities. Figure 5 visualizes carrier-dependent color responses across three examples, although the heatmaps may also reflect geometry, shading, and alignment and should not be interpreted as latent K or a pure color factor. Appendix F provides an additional decoded control grid.

Reading Carrier-Dependent Responses. Figure 5 provides spatial information that is obscured by global averages. Even under the same color reference, the two fits can induce different image-space changes because each preserves a different carrier. The response maps therefore indicate where the two edits agree or diverge, rather than which output better matches the target palette.

6 Discussion and Limitations

Scope and remaining comparisons. The quantitative subset covers one content object; broader content coverage, repeated seeds, and matched external baselines remain necessary. Direct-color assignment, shared-residual transfer, and geometry-fixed Gaussian-color optimization should be compared at matched color attainment to test the carrier-conditioning choice. For continuous control, strict-interior evaluation must be separated from the changed joint corner and should measure cross-attribute drift, multiview preservation, and monotonicity. Appendix E consolidates the evidence coverage and missing comparisons.

Practical limitations. Axis-dependent color allocation, soft geometry penalties, and per-instance fitting remain limitations. Fixed sparse support does not guarantee unchanged decoded geometry, and two latent coordinates do not establish exact disentanglement. Frozen weights also do not mean optimization-free authoring: texture sampling uses 80 steps and each color endpoint uses 30 fitting iterations. Intermediate states reuse endpoints but still require decoding and export; no aggregate runtime benchmark is reported.

7 Conclusion

We connect carrier-relative 3D color fitting with interaction-aware two-attribute morphing. Separate fitting around the source and textured carriers provides four aligned states; their color-displacement difference defines a texture-dependent color direction while preserving all four latent endpoints. The formulation operates without updating pretrained weights. The recorded experiments support inspectable endpoint generation and interaction effects, while the stated evaluation gaps delimit claims about comparative quality and independently controlled decoded attributes.

AI USE STATEMENT

For this manuscript revision, a generative AI assistant was used to suggest the organization of the main text and appendix, edit prose and captions, check internal consistency, and prepare and compile the LaTeX layout. This assistance did not include running new 3D-generation experiments or producing new numerical results. The authors retain responsibility for the final submitted text, claims, citations, and artifacts.

REPRODUCIBILITY STATEMENT

Section 3 defines the shared-support representation, carrier-conditioned objective, and latent control surface. Section 4.1 distinguishes generation coverage from the quantitative evaluation subset; Sections 4.2–5 specify the recorded geometry and interaction diagnostics. Appendix B retains the available preprocessing equations, initialization procedure, and numerical settings, while Appendix D separates unmatched external results from within-pipeline evidence. The documented implementation and evaluation limitations should be addressed when releasing executable code and complete experiment configurations.

REFERENCES

Aishwarya Agarwal, Srikrishna Karanam, and Balaji Vasan Srinivasan. Training-free color-style disentanglement for constrained text-to-image synthesis, 2024. URL https://arxiv.org/abs/2409.02429. arXiv:2409.02429.

Tomohiro Aizawa, Shigeru Kuriyama, and Chunzhi Gu. OrnaStyler: Ornament-aware latent editing for content-preserving 3D stylization, 2026. URL https://arxiv.org/abs/2608.29905. arXiv:2608.29905.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, 2021. URL https://arxiv.org/abs/2104.14294. arXiv:2104.14294.

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*, 2024. URL https://arxiv.org/abs/2312.09008.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A Neural Algorithm of Artistic Style, 2015. URL https://arxiv.org/abs/1508.06576. arXiv:1508.06576.

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*, 2024. URL https://arxiv.org/abs/2312.02133.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, 2020. URL https://arxiv.org/abs/2006.11239. arXiv:2006.11239.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large Reconstruction Model for Single Image to 3D. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.04400.

Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. URL https://arxiv.org/abs/1703.06868.

Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. StylizedNeRF: Consistent 3D scene stylization as stylized NeRF via 2D-3D mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2205.12183.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*, 2016. URL https://arxiv.org/abs/1603.08155.

Heewoo Jun and Alex Nichol. Shap-E: Generating Conditional 3D Implicit Functions, 2023. URL https://arxiv.org/abs/2305.02463. arXiv:2305.02463.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering, 2023. URL https://arxiv.org/abs/2308.04079. arXiv:2308.04079.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Wenzhao Li, Tianhao Wu, Fangcheng Zhong, and Cengiz Oztireli. ARF-Plus: Controlling perceptual factors in artistic radiance fields for 3D scene stylization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. doi: 10.1109/WACV61041.2025.00230. URL https://arxiv.org/abs/2308.12452.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1705.08086.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-Resolution Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2211.10440.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, 2022. URL https://arxiv.org/abs/2210.02747. arXiv:2210.02747.

Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric Xing. StyleRF: Zero-shot 3D style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a. URL https://arxiv.org/abs/2303.10598.

Kunhao Liu, Fangneng Zhan, Muyu Xu, Christian Theobalt, Ling Shao, and Shijian Lu. StyleGaussian: Instant 3D style transfer with gaussian splatting. In *SIGGRAPH Asia 2024 Technical Communications*, 2024. URL https://arxiv.org/abs/2403.07807.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot One Image to 3D Object, 2023b. URL https://arxiv.org/abs/2303.11328. arXiv:2303.11328.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, 2022. URL https://arxiv.org/abs/2209.03003. arXiv:2209.03003.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2003.08934.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A System for Generating 3D Point Clouds from Complex Prompts, 2022. URL https://arxiv.org/abs/2212.08751. arXiv:2212.08751.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy 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 Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2023. URL https://arxiv.org/abs/2304.07193. arXiv:2304.07193.

Dae Young Park and Kwang Hee Lee. Arbitrary Style Transfer with Style-Attentional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1812.02342.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion, 2022. URL https://arxiv.org/abs/2209.14988. arXiv:2209.14988.

Jiang Qin, Senmao Li, Alexandra Gomez-Villa, Shiqi Yang, Yaxing Wang, Kai Wang, and Joost van de Weijer. Free-lunch color-texture disentanglement for stylized image generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.14275.

Zefan Qu, Zhenwei Wang, Haoyuan Wang, Ke Xu, Gerhard Hancke, and Rynson W. H. Lau. StyleSculptor: Zero-shot style-controllable 3D asset generation with texture-geometry dual guidance, 2025. URL https://arxiv.org/abs/2509.13301. arXiv:2509.13301. SIGGRAPH Asia 2025, as reported by the authors.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2112.10752.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

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*, 2026. URL https://arxiv.org/abs/2601.00204.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. InstantStyle: Free Lunch towards Style-Preserving in Text-to-Image Generation, 2024. URL https://arxiv.org/abs/2404.02733. arXiv:2404.02733.

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*, 2025. URL https://arxiv.org/abs/2412.01506.

Zhiyu Xie, Yuqing Zhang, Xiangjun Tang, Yiqian Wu, Dehan Chen, Gongsheng Li, and Xiaogang Jin. StyleTex: Style image-guided texture generation for 3D models. *ACM Transactions on Graphics*, 43(6), 2024. doi: 10.1145/3687931. URL https://xzyw7.github.io/StyleTex/.

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, 2024. URL https://arxiv.org/abs/2404.07191. arXiv:2404.07191.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, 2023. URL https://arxiv.org/abs/2308.06721. arXiv:2308.06721.

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*, 2022. URL https://arxiv.org/abs/2206.06360.

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*, 2018. URL https://arxiv.org/abs/1801.03924.

A Extended Related Work

A.1 Image Stylization and Attribute Disentanglement

Neural style transfer represents content and style through different statistics of visual features. Gatys et al. (Gatys et al., 2015) formulate their recombination as image optimization, while perceptual-loss networks (Johnson et al., 2016) enable feed-forward image transformation. Adaptive instance normalization (Huang & Belongie, 2017) and whitening and coloring transforms (Li et al., 2017) transfer reference statistics without restricting inference to a single style. Style-attentional networks (Park & Lee, 2019) additionally model content–style feature interactions. These approaches provide foundations for feature-level transfer, but separating content from style does not automatically separate the texture and color components within style.

Diffusion models (Ho et al., 2020; Song et al., 2021) and latent diffusion (Rombach et al., 2022) provide another setting for reference-conditioned synthesis. IP-Adapter (Ye et al., 2023) introduces image prompting through decoupled text/image cross-attention. Style Injection (Chung et al., 2024) adapts diffusion inference for reference stylization, and StyleAligned (Hertz et al., 2024) shares attention to coordinate style across generated images. InstantStyle (Wang et al., 2024) investigates content–style separation and selective feature injection. These mechanisms address how a reference influences generation; their separation of content, image prompts, or style should not be conflated with guaranteed independence of color and texture in a decoded 3D asset.

Attribute-specific methods more directly inform our problem. Agarwal et al. (Agarwal et al., 2024) investigate separate color and style control using feature transformations and LAB-space factorization. SADis (Qin et al., 2025) studies separate texture and color references through color–grayscale embedding differences and regularized whitening and coloring transforms. We inherit the principle of paired attribute conditioning rather than claim color differencing as a new operation. Our focus is its realization in a shared 3D latent representation: paired differences initialize an edit, while carrier-relative decoder constraints address changes to geometry and local appearance that feature-space separation alone does not rule out.

A.2 Reference-Guided 3D Appearance Stylization

Neural radiance fields (Mildenhall et al., 2020) and 3D Gaussian splatting (Kerbl et al., 2023) provide representations in which appearance can be edited and rendered across viewpoints. Stylized-NeRF (Huang et al., 2022) couples 2D stylization and a radiance-field representation through mutual learning. ARF (Zhang et al., 2022) optimizes artistic radiance fields using reference features, while StyleRF (Liu et al., 2023a) enables zero-shot radiance-field style transfer. These methods address stylization in a shared scene representation rather than independently transforming each rendered view.

Perceptual control within 3D stylization is also established prior art. ARF-Plus (Li et al., 2025) provides controls including color preservation, pattern scale, and spatially selective stylization. StyleGaussian (Liu et al., 2024) transfers embedded features and decodes them into stylized Gaussian colors. StyleTex (Xie et al., 2024) combines style and geometry guidance for texture generation on a given mesh. Their input assumptions and editable representations differ, but they show that 3D appearance need not be treated as an indivisible factor.

Our setting retains a structured generative latent as the editable state and applies color fitting relative to an already textured carrier. This distinction does not imply that latent editing is inherently preferable to appearance-space editing: modifying only Gaussian colors can keep geometry fixed by construction. Direct color assignment and geometry-fixed Gaussian-color optimization are therefore essential comparisons. The research question is whether carrier-conditioned latent fitting offers a useful preservation–attainment trade-off while retaining a common state for subsequent texture–color morphing.

A.3 Structured-Latent Generation, Editing, and Morphing

Generative 3D modeling provides complementary ways to obtain an editable object state. DreamFusion (Poole et al., 2022) and Magic3D (Lin et al., 2023) use pretrained 2D diffusion guidance

for 3D optimization. Point-E (Nichol et al., 2022) generates point clouds, whereas Shap-E (Jun & Nichol, 2023) generates conditional implicit representations. Image-conditioned approaches include viewpoint-controlled synthesis in Zero-1-to-3 (Liu et al., 2023b), single-image reconstruction with LRM (Hong et al., 2024), and the multiview-generation/reconstruction pipeline of InstantMesh (Xu et al., 2024). These approaches concern object generation or reconstruction; they do not by their representation choice alone define separate texture and color controls.

TRELLIS (Xiang et al., 2025) supplies the structured latent representation and pretrained flow models used in our framework. Rectified flow (Liu et al., 2022) and flow matching (Lipman et al., 2022) provide the broader vector-field modeling context; our contribution is not a new flow-training objective. Likewise, self-supervised visual representations such as DINO and DINOv2 (Caron et al., 2021; Oquab et al., 2023) provide feature foundations for image conditioning and regional reference grouping, not semantic-part correspondence guarantees. Within generative stylization, StyleSculptor (Qu et al., 2025) uses cross-3D attention and style-related feature selection for texture and geometry control, making it a close comparison. OrnaStyler (Aizawa et al., 2026) studies ornament-aware, content-preserving latent editing. We distinguish our carrier-relative color formulation from the broader capabilities of staged or training-free 3D editing.

Structured latents also support transitions between objects. MorphAny3D (Sun et al., 2026) blends source and target latent features within attention mechanisms to construct morphing sequences. Our control domain instead spans texture and color strengths on one shared sparse support. Four attribute-specific endpoints supply two carrier-conditioned color displacements, whose difference is retained in a bilinear surface. The distinction is the organization of these edits into an interpretable two-attribute domain, not the invention of bilinear interpolation or a claim that independently executed edit operators commute.

B Implementation Details

B.1 Detail-Oriented Reference Preparation

The main text defines the frequency-selective transform and distinguishes grouping descriptors from attention tokens. Here we expand the preprocessing steps and enlarge the reference-preparation part of the supplied vector diagram. The grouping features route reference tokens; they are not substituted for the original tokens in the key/value projections.

Frequency-selective preparation. Equations (1)–(2) in the main text specify the complete high-pass decomposition and neutral-reference reconstruction. H_Y retains local luminance contrast, whereas H_χ subtracts the luminance contribution from the RGB residual. The two normalized residuals are scaled separately before clipping. Thus the construction does not simply discard all chromatic information: local color variations can remain part of the transferred texture.

Token and descriptor pathways. Figure B.1 expands foreground extraction, cropping, resizing, and reference grouping. Patch descriptors exclude prefix tokens and are aligned with the luminance, detail, and coordinate cues on the same image patch grid. The descriptor concatenation defines groups; it does not replace the original reference tokens used for attention keys and values. The spatial cues refer to the reference image, not known correspondences with semantic 3D parts. The recorded settings and remaining implementation specifications are collected in Sec. B.3.

B.2 Color-Residual Initialization

These equations specify the shared initialization used before the two independent fits in Sec. 3.3. They do not replace carrier-relative optimization or imply globally disentangled color coordinates.

Color Residual Estimation. Following paired color–grayscale conditioning (Qin et al., 2025), we sample auxiliary color, grayscale, and neutral latents F_{col}, $F_{\mathrm{col},g}$, and $F_{\mathrm{col},n}$ using matched noise, together with a grayscale source latent $F_{s,g}$. We estimate an initial color displacement as

$$\Delta C_{\mathrm{init}} = \mathcal{V}\left(F_{\mathrm{col}} - F_{\mathrm{col},g}, F_{\mathrm{col},g} - F_{\mathrm{col},n}\right) - \left(F_{00} - F_{s,g}\right), \tag{B.1}$$

where $\mathcal{V}$ suppresses variation associated with the gray-to-neutral change, reducing luminance leakage. Since F_{00} is not exactly paired with $F_{s,g}$ under the same sampling trajectory, ΔC_{init} provides an initialization rather than an exact color factor.

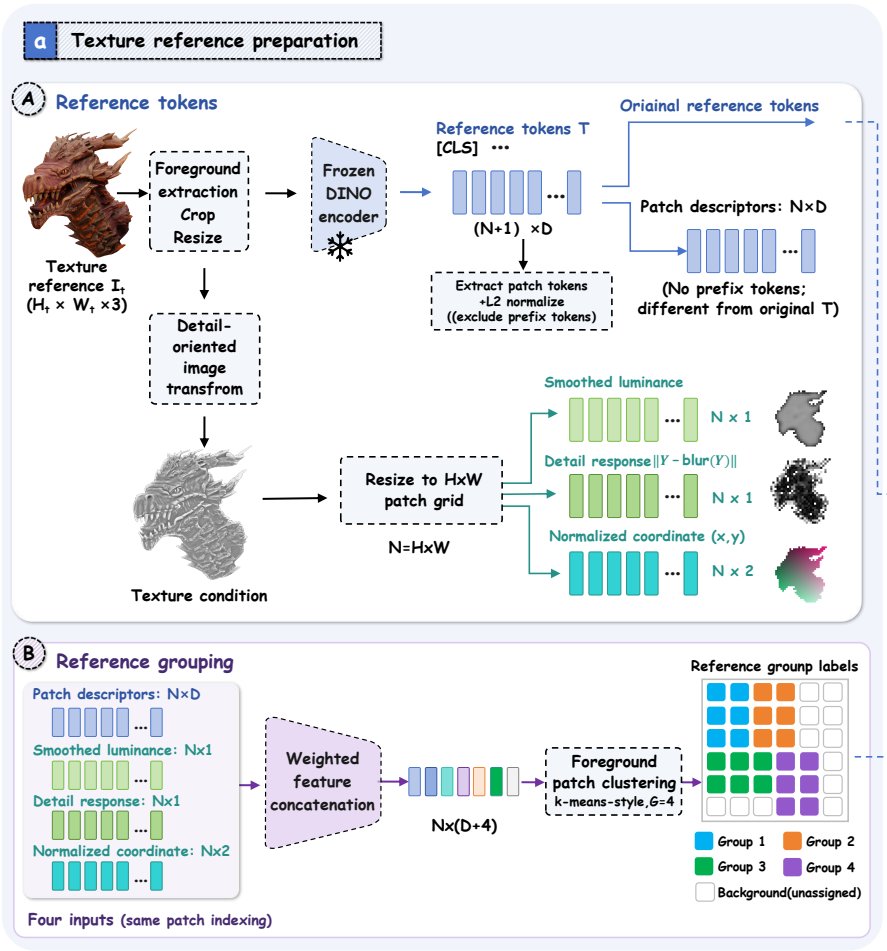


Figure B.1: **Enlarged reference preparation and grouping.** This is panel a of the original pipeline, shown separately to keep its token and descriptor pathways readable. Patch descriptors exclude prefix tokens; the grouping inputs combine descriptors, smoothed luminance, detail response, and coordinates. Grouping does not imply semantic-part correspondence.

Texture-Aware Residual Refinement. The initial color displacement may still contain texture-related components. We compute the observed texture displacement as $\Delta T = F_{10} - F_{00}$ and extract its high-frequency component as $T_H = \Delta T - \mathcal{S}_{r_t}(\Delta T)$, where $\mathcal{S}_{r_t}$ denotes local smoothing over the sparse SLAT support. Similarly, we smooth the initial color displacement as $\Delta C = \mathcal{S}_{r_c}(\Delta C_{\text{init}})$.

We then suppress the component of ΔC aligned with the observed texture variation:

$$\widehat{\Delta C} = \Delta C - \text{Proj}_{T_H}(\Delta C). \tag{B.2}$$

Here, $\text{Proj}_{T_H}(\Delta C)$ denotes the projection of the color displacement onto the high-frequency texture direction. In practice, the projection is performed channel-wise, followed by normalization and RMS rescaling; implementation details are provided in the Appendix.

The refined color displacement is used to initialize the two color-editing branches:

$$F_{01}^{(0)} = F_{00} + \gamma\widehat{\Delta C}, \qquad F_{11}^{(0)} = F_{10} + \gamma\widehat{\Delta C}, \tag{B.3}$$

where γ controls the initialization strength.

B.3 Recorded Settings and Reporting Scope

Available implementation description. Regional guidance combines reference appearance and spatial descriptors with an attention response from grouped reference tokens. Sparse smoothing and

Table B.1: Settings used for the completed 120-triplet benchmark. Values are retained from the recorded configuration; see the reporting scope below.

Setting	Value
Backbone / output	TRELLIS image-large / Gaussian
Sparse / auxiliary / texture-carrier steps	8 / 8 / 80
Color optimizer / iterations	Adam / 30
Learning rate / gradient clip	0.015 / 1.0
Initialization γ / target strength	0.25 / 1.0
$(\lambda_{\mathrm{dc}}, \lambda_{\mathrm{ch}}, \lambda_{\mathrm{val}})$	$(6, 0.5, 2)$
$(\lambda_{\mathrm{det}}, \lambda_{\mathrm{geo}}, \lambda_z)$	$(3, 2, 0.03)$
$(\omega_p, \omega_s, \omega_q, \omega_o)$	$(2, 1, 0.35, 0.75)$
Decoder-loss voxel resolution	16^3
Geometry-preservation mechanism	Soft decoded-field penalties
Recorded device	NVIDIA RTX 6000 Ada Generation (48 GB)

channel-wise projection refine the color residual, followed by normalization and RMS rescaling. Decoder-space color targets adapt palette distributions to the current carrier's luminance and shading; only latent features are fitted.

Details requiring executable specification. The retained manuscript does not fully specify the normalization operators $\mathcal{N}_Y, \mathcal{N}_\chi$, the implementation of $\mathcal{V}$, the sparse smoothing/projection neighborhood, or the palette extraction and spatial-assignment routine. The recorded configuration also lists loss subweights $(\lambda_{\mathrm{dc}}, \lambda_{\mathrm{ch}}, \lambda_{\mathrm{val}}, \lambda_{\mathrm{det}}, \lambda_{\mathrm{geo}})$ and $(\omega_p, \omega_s, \omega_q, \omega_o)$ without complete subterm definitions. These values are retained for traceability, not presented as a complete executable specification. The evaluated channel-overlap switch belongs to the recorded texture-carrier implementation; it should not be read as evidence for a globally disjoint partition of texture and color channels.

C Latent Surface Identities

The interaction construction in Sec. 3.4 can be expanded without assumptions on the nonlinear decoder:

$$\begin{aligned} F(\alpha,\beta) &= F_{00} + \alpha(F_{10} - F_{00}) + \beta(F_{01} - F_{00}) + \alpha\beta(F_{11} - F_{10} - F_{01} + F_{00}) \\ &= (1-\alpha)(1-\beta)F_{00} + \alpha(1-\beta)F_{10} + (1-\alpha)\beta F_{01} + \alpha\beta F_{11}. \end{aligned} \tag{C.1}$$

Substitution at $(\alpha, \beta) \in \{0,1\}^2$ yields the four fitted endpoints. The latent derivatives are

$$\frac{\partial F}{\partial \beta} = \Delta C_s + \alpha K, \qquad \frac{\partial F}{\partial \alpha} = \Delta T + \beta K, \qquad \frac{\partial^2 F}{\partial \alpha\, \partial \beta} = K. \tag{C.2}$$

Thus K is the mixed finite difference of the four anchors and the mixed derivative of their bilinear surface. It is not a measured semantic interaction in decoded image space. The additive comparator omits $\alpha\beta K$, so the two latent surfaces differ by exactly that amount and agree wherever $\alpha = 0$ or $\beta = 0$. At $(1,1)$ their difference is K.

Decoded interpretation. A deterministic decoder evaluated at an identical fitted corner gives the corresponding fitted output. Interior changes are mediated by the decoder: even a smooth latent path need not be perceptually monotone or preserve geometry. The identities above therefore justify endpoint recovery and coordinate-dependent latent directions, not independent editing operators, perceptual path independence, or a quality advantage over other interpolation schemes.

D Contextual Cross-Method Results

Table 1 in Sec. 4.2 reports the complete recorded cross-method reference-agreement results. This appendix clarifies their interpretation and retains the external geometry rows. Different evaluation samples and incompletely matched implementation details prevent a controlled cross-method ranking; no cross-method best-value highlighting is used.

The SSIM values in Table 1 retain their original $\times 10^3$ display scale; our displayed value 81.84 corresponds to an unscaled SSIM of 0.08184. VLM-C and VLM-T are reproduced as recorded,

but the evaluator identity, prompt, scoring scale, and aggregation protocol are not specified in the available manuscript. They therefore do not support a calibrated quantitative claim.

Table D.1: Contextual geometry-preservation distances ($\times 10^{-2}$; lower is better). Our three edited endpoints are compared with their generated F_{00} meshes on 10 triplets using 100,000 sampled points per mesh. External rows use a different evaluation set and incompletely matched implementations; no cross-method ranking is implied.

Method	Chamfer distance ↓	Hausdorff distance ↓
MorphAny3D	198.52	44.50
SADis	18.17	41.59
StyleSculptor	17.91	8.38
Ours–Texture (F_{10})	1.38	6.56
Ours–Color (F_{01})	1.30	5.95
Ours–Joint (F_{11})	1.44	7.37

For our endpoint rows, Sec. 4.2 defines the unit-sphere normalization and point sampling. The original external row label "SADis" is retained verbatim here, whereas the visual comparison uses "SADis+TRELLIS"; without a matching protocol record these labels should not be assumed interchangeable. The relatively large external Chamfer values should be checked for distance definition, normalization, and units before any comparison is made.

E Evaluation Limitations and Required Comparisons

Table E.1 summarizes the distinct scopes of generation, quantitative evaluation, and control diagnostics.

Table E.1: **Evidence coverage.** Generation success, quantitative evaluation, and control-grid diagnostics have different scopes and must not be conflated.

Evidence	Scope	Evaluated output
Generation / export	120 balanced triplets	All four endpoints (480 GLBs)
Reference agreement	10 triplets, one content	Joint endpoint F_{11}
Geometry distances	Same 10 triplets	F_{10}, F_{01}, F_{11} vs. F_{00}
Implementation ablations	Same 10 triplets	Joint endpoint F_{11}
Interaction ablation	One fireplace case	Two 7×7 decoded grids
Additional control grid	One bull case	One 7×7 decoded grid

The endpoint benchmark records 120 completed triplets and 480 exported GLBs. Its ten-triplet quantitative subset and one-case interaction comparison do not establish broad comparative superiority or exact disentanglement. The 25 strict interior points in the fireplace case differ by 13.07 ± 7.35 foreground RGB levels after removing K, while the joint corner differs by 36.40. These are changes between variants, not improvements against reference targets.

The central carrier-conditioning hypothesis requires shared-residual, direct-color, and geometry-fixed Gaussian-color baselines at matched color attainment. Their corresponding images and scores are not available for the reported split, so no placeholder values are provided. Direct ablations should isolate carrier-relative fitting, residual refinement, and regional guidance, rather than infer their effects from the available texture-carrier switches.

For continuous control, a reference-based evaluation should separately report strict-interior performance and the fitted joint endpoint, because the additive comparator changes the latter by construction. Cross-axis drift, target attainment, perceptual monotonicity, and multiview depth/silhouette preservation need evaluation over more objects and views. Repeated seeds and runtime measurements are also needed to assess robustness and practical editing cost. These gaps are limitations of the available evidence, not additional experiments claimed by this manuscript.

F Additional Qualitative Results

Carrier-conditioned endpoint examples. Figure F.1 shows additional endpoint examples for the separate carrier-relative fits in Sec. 3.3.

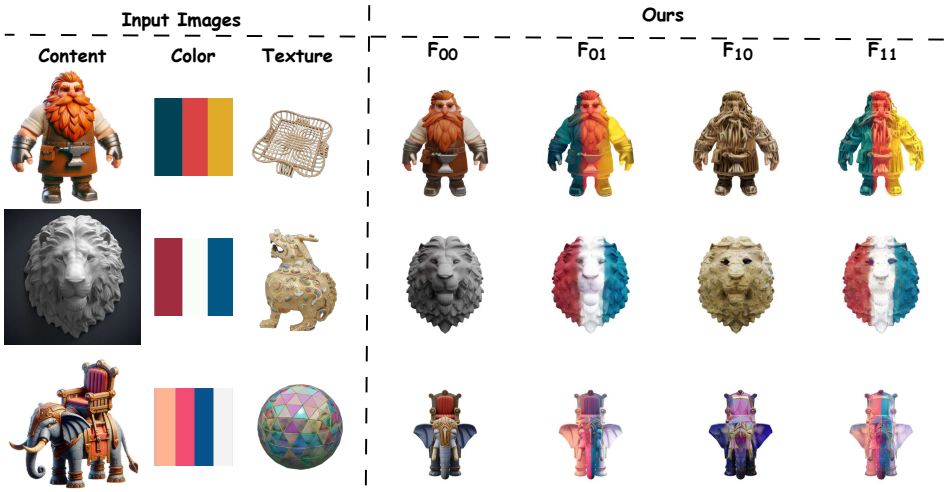


Figure F.1: **Four aligned endpoints from separate carrier-relative fits.** A character, an excavator, and a lion-head object are shown with their content, texture, and palette references. Endpoint order is F_{00}, F_{10}, F_{01}, and F_{11}, as labeled in the supplied asset.

Continuous control example. Figure F.2 provides an additional 7×7 example using a bull character, woven-chair texture, and five-color palette. References are shown at the left. Color increases to the right and texture downward; all 49 cells are Gaussian-decoder renders of latent states sampled at increments of $1/6$, not image blends. This case is not included in the single-case numerical interaction ablation. It illustrates a second object's control range but does not test monotonicity or reference attainment quantitatively.

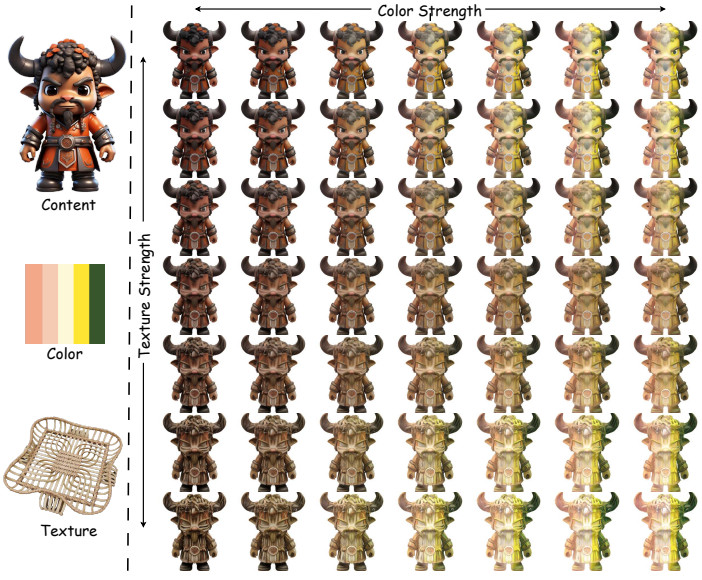


Figure F.2: **Additional two-axis control example.** The bull grid samples texture and color coordinates in steps of $1/6$. Inputs occupy the left strip. The figure is a qualitative example, not a second quantitative interaction ablation.