BindStyle: Single-image dynamic Gaussian stylization via mesh-bound style propagation

Anonymous authors
Paper under double-blind review

ABSTRACT

Stylizing a dynamic 3D scene from a single image requires colors, textures, and geometric changes to remain attached to the same scene elements as they move. Independent frame-wise edits do not enforce this association, while appearance-only edits leave spatial support unchanged. We propose BindStyle, a mesh-bound Gaussian framework that learns time-shared appearance and local geometry from a style reference. Given a pretrained Gaussian scene and a corresponding mesh sequence, the binding transports the learned attributes through the supplied motion while mesh vertices and attachment identities remain fixed. Color changes and geometric changes can explain the same rendered discrepancy, so BindStyle controls which parameters each supervisory signal updates. Local VGG feature matching guides appearance through spherical-harmonic coefficients, whereas reference-conditioned video diffusion guides local positions, scales, and rotations using mesh depth and temporal context. The reference schedule first adapts appearance, then enables geometry while continuing appearance updates. This gives geometric adaptation an appearance-informed starting point and lets appearance adjust as Gaussian support changes. The resulting model renders the supplied motion without per-frame style optimization or diffusion inference. Multi-style comparisons and matched-teacher ablations support improved artistic appearance and temporal consistency with parameter-routed supervision. An anonymous source snapshot is available at https://anonymous.4open.science/r/BindStyle-23C0/.

1 INTRODUCTION

Dynamic 3D stylization must keep appearance and geometric edits attached to the same scene elements across motion and viewpoints. Fig. 1 contrasts independent frame-wise edits and appearance-only transfer with BindStyle's shared appearance and local-shape editing under known motion. Gaussian splatting exposes appearance and spatial support through differentiable rendering (Kerbl et al., 2023), while mesh binding associates these primitives with a moving surface (Ma et al., 2025).

Given a pretrained Gaussian reconstruction, a corresponding mesh sequence, calibrated cameras, and one static style image, we learn style while keeping motion fixed. Existing methods transfer reference appearance to reconstructed 4D scenes (Liang et al., 2026; Chen et al., 2026). We study shared appearance and editable mesh-local Gaussian support under supplied motion. The meshes specify how scene elements move, while the image specifies their desired style; neither supplies ground-truth stylized geometry.

BindStyle stores these attributes directly in a shared set of Gaussian parameters, as shown in Fig. 2. Each primitive retains its face attachment, while its spherical-harmonic (SH) coefficients, mesh-local position, scale, and rotation are optimized. The MaGS binding operator (Ma et al., 2025) maps the same local parameters to frame-dependent world-space primitives. Rendered temporal windows supervise this representation during optimization. At inference, the given mesh sequence transports the learned attributes without per-frame style prediction or optimization. Thus style learning operates on persistent scene elements, whereas motion propagation reuses the existing binding.

An image-space discrepancy can be reduced by changing color, spatial support, or visibility, making its attribution to Gaussian parameters ambiguous. SH-only editing excludes geometric expression,

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.

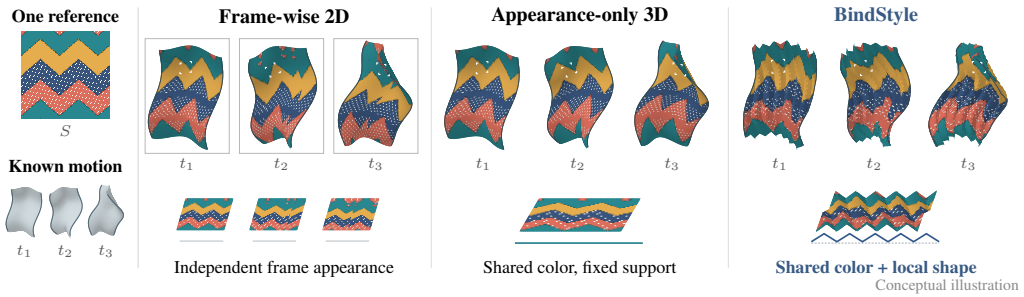

Figure 1: **Persistent appearance and local shape under motion.** Frame-wise editing lacks shared 3D attributes; appearance-only editing fixes local support. BindStyle shares both appearance and surface-relative Gaussian support through the supplied motion. White marks identify one surface patch. Insets depict Gaussian support changes with fixed mesh vertices. This is a conceptual illustration.

while applying every objective to every parameter permits competing explanations of the same reference cue. We address this ambiguity through explicit supervisory routing. Nearest-neighbor VGG feature matching updates SH appearance. A reference-conditioned video diffusion teacher updates local Gaussian geometry, conditioned on the style image, mesh-derived depth, and a short temporal window. The assignment is imposed by the optimization rule; it does not assume that either teacher internally separates appearance from shape.

The reference schedule first optimizes appearance, then enables geometry while continuing appearance optimization. It is intended to limit geometric compensation for the initial appearance mismatch, rather than guarantee the best final color matching. Both branches use the same rendered state but update disjoint parameter groups. Binding supplies correspondence, depth supplies structure, and temporal modules supply a video prior; their effects require empirical evaluation.

Our contributions are twofold.

- To the best of our knowledge, we are the first to jointly learn time-shared Gaussian appearance and mesh-local geometry from a single style image and transport both through a prescribed mesh sequence. This formulation separates style learning from known-motion propagation without per-frame style prediction or relearning motion.
- A parameter-routed optimization scheme that assigns local feature matching to appearance and source-referenced video diffusion guidance to geometry, with explicit gradient ownership and an analysis of temporal context and optimization order.

2 Related Work

Dynamic scene representations. NeRF represents density and view-dependent radiance continuously (Mildenhall et al., 2020). D-NeRF models deformation, whereas HexPlane factorizes space–time features (Pumarola et al., 2021; Cao & Johnson, 2023). Gaussian representations support explicit optimization of appearance and spatial extent (Kerbl et al., 2023); 4D-GS predicts time-dependent deformation, while Dynamic 3D Gaussians track persistent primitives (Wu et al., 2024; Luiten et al., 2024). SuGaR connects Gaussians to reconstructed surfaces (Guédon & Lepetit, 2024), and MaGS uses mesh attachment for dynamic reconstruction and simulation (Ma et al., 2025). We inherit this attachment to transport shared style parameters through known motion rather than learning another motion field.

Image, video, and 3D stylization. Neural style transfer uses feature statistics, perceptual objectives, normalization, or attention (Gatys et al., 2016; Johnson et al., 2016; Huang & Belongie, 2017; Liu et al., 2021). Video methods address temporal stability (Ruder et al., 2016; Huang et al., 2017; Gupta et al., 2017); StyleID and StyleMaster use diffusion for image and video stylization (Chung et al., 2024; Ye et al., 2025). For 3D scenes, StylizedNeRF couples 2D and 3D learning, ARF uses nearest-neighbor matching, and StyleRF transforms radiance-field features (Huang

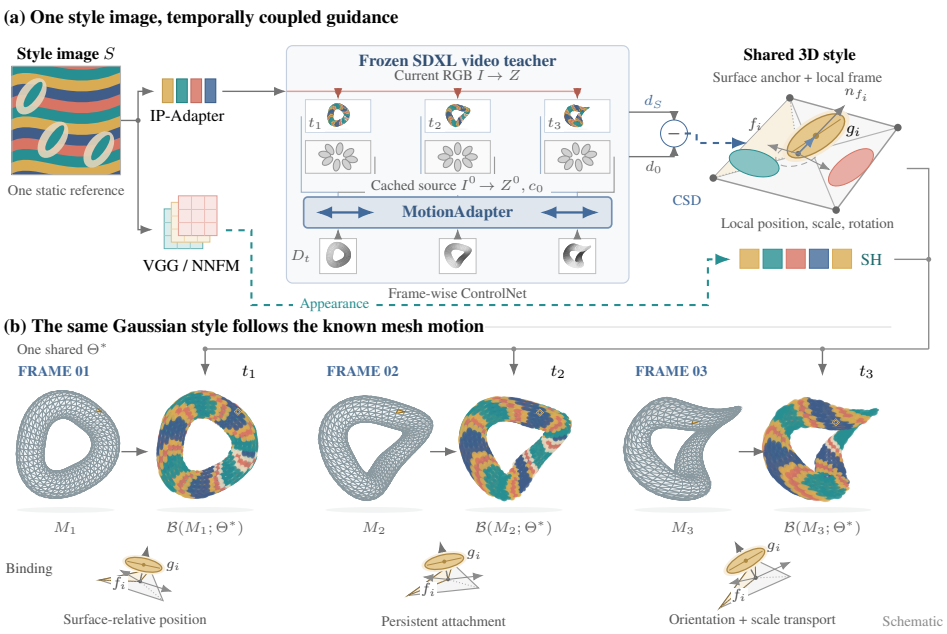

Figure 2: **Shared style learning and motion propagation.** (a) The frozen teacher compares style-conditioned current renders with cached initial renders conditioned on their source images. Both branches share depth, timestep, and noise; their guidance difference $d_S - d_0$ supplies video CSD. Dashed paths update local Gaussian support with CSD and SH appearance with VGG matching through RGB renders. (b) A persistent face attachment transports the same Gaussian's center, orientation, and scale through the supplied mesh motion. Gold markers track one attachment. The schematic distinguishes the teacher's temporal prior from known scene motion.

et al., 2022; Zhang et al., 2022; Liu et al., 2023). StyleGaussian learns Gaussian feature transformation and decoding (Liu et al., 2024). G-Style, ReGS, SGSST, and StylizedGS explore scene-specific Gaussian stylization (Kovács et al., 2024; Mei et al., 2024; Galerne et al., 2025; Zhang et al., 2025), while Stylos uses a feed-forward architecture (Liu et al., 2026). Dynamic methods include feature-space transfer in 4DStyleGaussian and ZDySS, and space–time-aware appearance control in Style4D (Liang et al., 2026; Saroha et al., 2025; Chen et al., 2026). Dynamic mesh stylization also edits vertex positions with temporal control (Haetinger et al., 2024). We instead retain the input meshes and jointly learn shared Gaussian appearance and mesh-local support, with separate supervision for each parameter group.

Diffusion guidance and conditioning. Denoising diffusion and latent diffusion provide the generative basis of our SDXL teacher (Ho et al., 2020; Rombach et al., 2022; Podell et al., 2024). DreamFusion and DreamGaussian distill image priors into 3D representations (Poole et al., 2023; Tang et al., 2024); ProlificDreamer studies variational score distillation (Wang et al., 2023). CSD emphasizes classifier guidance, whereas DDS uses source-referenced denoising differences (Yu et al., 2024; Hertz et al., 2023). Instruct-NeRF2NeRF instead alternates image editing and scene reconstruction (Haque et al., 2023). FantasyStyle, DiffStyle3D, and SIC3D adapt diffusion evidence to Gaussian stylization (Yang et al., 2026b;a; He et al., 2026); GeoStyle edits mesh geometry (Choi et al., 2026). We retain the supplied meshes and use a source-referenced difference of classifier-guidance directions, not the original SDS, CSD, or DDS estimator. CLIP image features and IP-Adapter provide reference conditioning (Radford et al., 2021; Ye et al., 2023), ControlNet supplies structure (Zhang et al., 2023), and AnimateDiff supplies temporal context (Guo et al., 2024). These frozen components guide Gaussian support; they do not provide explicit 3D shape targets or replace mesh correspondence.

3 METHOD

3.1 PROBLEM FORMULATION

Given a style image S, we stylize a pretrained mesh-bound Gaussian scene observed by content images $\{J_t\}$, cameras $\{C_t\}$, and foreground masks $\{m_t\}$. A mesh sequence $M_t = (V_t, F)$ with shared topology and face correspondence specifies the known motion. The content images and meshes describe the source scene; neither is a target for stylized appearance or shape.

The model contains N Gaussian primitives with fixed face assignments f_i and opacities. We optimize one shared parameter set $\Theta = (A, G)$, partitioned into appearance and local geometry:

$$A = \{c_i^{\mathrm{DC}}, c_i^{\mathrm{rest}}\}_{i=1}^{N}, \qquad G = \{u_i, v_i, w_i, \ell_i, q_i\}_{i=1}^{N}, \tag{1}$$

where (u_i, v_i, w_i) are mesh-local coordinates, $\ell_i \in \mathbb{R}^3$ is log-scale, and $q_i \in \mathbb{R}^4$ is a rotation quaternion. The objective is to obtain Θ^* whose renders express S across the supplied motion and camera views while preserving recognizable content. These are direct parameter updates, not outputs of a style-prediction network or additive style field. Meshes, face assignments, and opacities remain fixed. The pretrained deformation network is excluded from the stylization forward pass.

The output is an edited Gaussian model that renders with the same mesh sequence. Throughout the paper, geometric stylization denotes changes to Gaussian support and its rendered effects, not edits to the mesh vertices.

3.2 MESH-BOUND STYLE PROPAGATION

Shared representation. The learned appearance and local geometry belong to persistent Gaussian primitives. Motion changes their world-space realization, not their stored style parameters. Let $\mathcal{B}$ denote mesh binding and $\mathcal{R}$ Gaussian rendering. For time t and camera $C_{t,v}$, the dynamic forward map is

$$\hat{I}_{t,v}(\Theta) = \mathcal{R}\big(\mathcal{B}(M_t; \Theta), C_{t,v}\big). \tag{2}$$

The same Θ is used at every time and view. In the training notation below, $I_t = \hat{I}_t(\Theta)$ denotes the render at the supplied camera C_t.

Transporting local geometry. We use the inherited MaGS binding (Ma et al., 2025). For the three vertices $v_{f_i,t}^{0:2}$ of the attachment face and its supplied normal $n_{f_i,t}$, the Gaussian center is

$$x_{i,t} = u_i v_{f_i,t}^0 + v_i v_{f_i,t}^1 + (1 - u_i - v_i) v_{f_i,t}^2 + w_i n_{f_i,t}. \tag{3}$$

The normalized primitive rotation is composed with the face orientation, and local scale is multiplied by the face's scalar size factor. Binding therefore transports covariance as well as position. An anisotropic Gaussian rotates and changes size with its attachment face. Appendix A.1 describes the implementation. No additional motion field is predicted.

Learning versus propagation. During optimization, windows sampled across time update the same A and G. During inference, Θ^* is fixed and Equation (2) transports its attributes through the supplied meshes. This shared parameterization preserves attribute identity at the primitive level. It does not enforce identical image appearance across time: visibility, footprint overlap, and view-dependent SH still influence the rendered sequence.

3.3 PARAMETER-ROUTED VISUAL SUPERVISION

We assign distinct supervisory signals to appearance and geometry. Local feature matching provides an explicit reference for appearance, while a conditioned video teacher supplies a sequence-level update direction for Gaussian support. Both objectives are evaluated on Gaussian RGB renders; their gradients are routed to different parameter blocks.

Appearance from local reference features. A frozen VGG16 network (Simonyan & Zisserman, 2015) extracts spatial features from the current renders and style image. NNFM (Zhang et al., 2022) matches each rendered feature to its nearest style feature without requiring pixel correspondence.

Let $a^l_{t,p}$ and s^l_r denote unit-normalized render and style features at layer $l \in \mathcal{L}$. For a window $\mathcal{W}$ of F_w frames and P_l render features per frame, the appearance objective is

$$\mathcal{L}_{\text{app}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \frac{1}{F_w P_l} \sum_{t \in \mathcal{W}} \sum_{p=1}^{P_l} \min_r \left(1 - \langle a^l_{t,p}, s^l_r \rangle\right). \tag{4}$$

Render features are extracted from a fixed foreground crop, and style features from the full reference. Layer-wise losses are averaged. This objective updates SH only; no Gram, AdaIN, or content reconstruction term is added. Appendix A.1 specifies feature extraction and matching.

Conditioned video teacher. The geometry branch combines three conditions within a frozen SDXL-based teacher (Podell et al., 2024). IP-Adapter supplies tokens from the style image (Ye et al., 2023). ControlNet receives frame-wise depth from the input meshes (Zhang et al., 2023). AnimateDiff motion modules couple frames within the rendered window (Guo et al., 2024). These provide a visual reference, structural conditioning, and a temporal prior, respectively. Unlike mesh binding, the temporal prior does not define the scene's motion.

Each window contains consecutive dataset frames with their associated meshes and cameras. Camera pose may vary within a window. Mesh depth is rasterized with nvdiffrast (Laine et al., 2020) and cached without gradients. ControlNet residuals are computed frame-wise and injected into the motion denoiser. The optimized image remains the Gaussian RGB render; mesh rasterization supplies only the fixed structural condition.

Source-referenced video guidance. Before optimization, the same forward map produces a fixed source sequence $I^0_t = \hat{I}_t(\Theta_0)$ with deformation inference disabled. These are initial Gaussian renders, not the photographs J_t, and are never refreshed. A frozen VAE encoder E maps current and source windows to $Z = E(I_{\mathcal{W}})$ and $Z^0 = E(I^0_{\mathcal{W}})$. Gradients are retained through the current encoding, whereas Z^0 is detached.

Current and source latents are perturbed using the same diffusion timestep τ and noise realization. Let d_S denote the style-conditioned classifier-guidance direction at the noisy current latent, and d_0 the source-image-conditioned direction at the noisy source latent. Both retain the same depth condition. We inject their difference through

$$g_Z = \text{clip}\left((1 - \bar{\alpha}_\tau)(d_S - d_0), -c, c\right), \qquad \mathcal{L}_{\text{geo}} = \frac{1}{BF_w} \langle \text{sg}(g_Z), Z \rangle. \tag{5}$$

Here B is the batch size, $\bar{\alpha}_\tau$ the cumulative signal coefficient, c the elementwise clipping threshold, and sg stop-gradient. Each direction is a scaled conditional-minus-unconditional noise prediction; Appendix A.3 specifies all four predictions. We refer to this estimator as source-referenced video CSD. Equation (5) injects $g_Z/(BF_w)$ into the current latent without differentiating the denoiser. It is a single-timestep update, and its signed surrogate value is not a perceptual quality measure.

Explicit gradient routing. At update k, both objectives use renders from the same parameter state (A_k, G_k). Their assigned gradients are

$$g_A = \lambda_A \partial_A \mathcal{L}_{\text{app}}(A_k, G_k), \qquad g_G = \lambda_G \partial_G \mathcal{L}_{\text{geo}}(A_k, G_k). \tag{6}$$

SH is held constant for g_G, and geometry is held constant for g_A. The geometry gradient propagates through the VAE input, Gaussian renderer, and mesh binding to G; it does not update the teacher or mesh. Unlike unrestricted minimization of a weighted loss sum, Equation (6) excludes $\partial_A \mathcal{L}_{\text{geo}}$ and $\partial_G \mathcal{L}_{\text{app}}$. Separate optimizers apply the two updates only after both derivatives have been computed. This separates direct supervisory responsibility, not the groups' effects on subsequent renders.

3.4 Surface-Relative Staged Optimization

Surface-relative editing. The known meshes define attachment identities, local frames, and motion correspondence. Geometry updates therefore act in a surface-relative coordinate system rather than independently changing world-space centers in each frame. Log-scales ensure positive sizes after exponentiation, and quaternions are normalized before composition with face orientation. These parameterizations preserve the form of a Gaussian, but do not impose a surface-distance constraint:

u_i, v_i are not projected onto a face simplex, and w_i is not hard-bounded. No additional surface or neighborhood penalty is used.

Mesh depth supplies structural evidence to the teacher rather than a loss equating Gaussian depth with the input surface. Accordingly, the original mesh acts as a motion and structure reference, not a prescribed stylized shape. Conditioning strength controls the influence of that reference; excessive offsets, sizes, or silhouette changes remain possible.

Reference optimization schedule. The first stage updates only A with Equation (4). At the geometry onset, local coordinates, scales, and rotations are enabled together. The second stage follows Equation (6): VGG continues to update appearance, while video guidance updates geometry. Both optimizers step at complete-window boundaries. Appearance-first staging provides an appearance-adapted starting point; simultaneous optimization instead enables both groups from the outset. Both schedules preserve the same gradient routing. Appendix C.2 compares their convergence behavior.

Appearance remains attached to the same primitives as their support changes, with continued VGG updates adapting it to the new render. Appendix A specifies the schedule and numerical safeguards. Inference requires only the edited Gaussians, meshes, and cameras; the visual teachers and deformer are absent.

4 Experiments

4.1 Experimental Setup

We evaluate on D-NeRF (Pumarola et al., 2021) and DG-Mesh (Liu et al., 2025) against StyleGaussian (Liu et al., 2024), StylizedGS (Zhang et al., 2025), 4DStyleGaussian (Liang et al., 2026), Style4D (Chen et al., 2026), and FantasyStyle (Yang et al., 2026b). Static methods process per-time snapshots; dynamic methods retain native representations. Consistency aggregates cover six selected scenes; quality comparisons use nine pairs spanning five scenes and three references, and routing ablations use three matched cases. Views, times, and references match within each comparison; BindStyle additionally uses supplied mesh correspondence and depth. Appendix B details the distinct evaluation sets and aggregation.

Evaluation Metrics. ArtScore (Chen et al., 2024) measures artistic appearance, ArtFID (Wright & Ommer, 2022) combines content and distribution distances, and CM Loss (Afifi et al., 2021) measures reference-color matching. Warped LPIPS (Zhang et al., 2018) and RMSE assess consistency on shared valid regions. Paired blind ratings complement these measures. Appendix B defines visibility, coverage, aggregation, and confidence intervals. Bold ranks quality/error scores, not coverage.

4.2 Comparison with Existing Methods

Reference-specific patterns distinguish style transfer from a palette change alone. Fig. 3 shows stained-glass boundaries and structured motifs, with recognizable poses and parts. The quality gains in Table 2 extend to all nine reported combinations: BindStyle improves ArtScore and color matching over both dynamic baselines. Paired ratings from 24 participants favor both style fit and content preservation, with positive confidence intervals against each comparator. Appendices B.5 and B.6 separate per-pair results, paired ratings, and ArtFID image-set evaluation.

View consistency tests whether patterns remain associated with scene elements as projection and visibility change. Common-mask evaluation in Table 1 preserves BindStyle's advantage at both baselines. All methods use the same visible support, so this comparison does not reward a smaller method-specific mask. Wide baselines remain harder as occlusions and projected footprints change.

Temporal consistency requires patterns to persist through pose changes. Shared-mask scores in Table 3 favor BindStyle over native dynamic methods at both gaps. Recurring motifs on the moving LEGO parts in Fig. 4 illustrate this propagation. Across view and time groups, common-mask LPIPS reductions relative to Style4D range from 8.1% to 15.6%, indicating that the advantage persists under identical evaluation support.

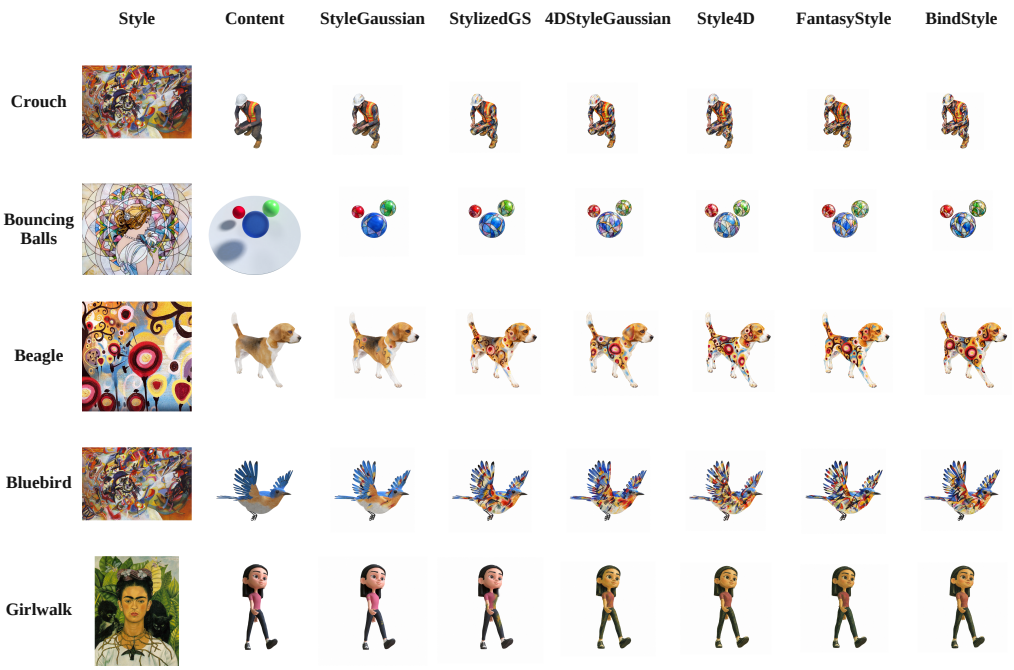

Figure 3: **Qualitative comparison.** Each row shows the reference, source, StyleGaussian, StylizedGS, 4DStyleGaussian, Style4D, FantasyStyle, and BindStyle. The examples compare reference-palette transfer, pattern coverage, and recognizable structure.

	Method-specific masks				Common masks			
	Narrow-baseline		Wide-baseline		Narrow-baseline		Wide-baseline	
Method	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓
StyleGaussian (2024)†	0.4539	0.0412	0.5788	0.1414	0.4612	0.0435	0.5894	0.1487
StylizedGS (2025)†	0.4098	0.0337	0.5225	0.1179	0.4175	0.0356	0.5331	0.1242
FantasyStyle (2026)†	0.3708	0.0266	0.4659	0.0964	0.3784	0.0283	0.4762	0.1021
4DStyleGaussian (2024/2026)	0.3450	0.0213	0.4127	0.0808	0.3526	0.0228	0.4235	0.0859
Style4D (2026)	0.3220	0.0175	0.3791	0.0694	0.3297	0.0189	0.3894	0.0741
Ours	**0.2963**	**0.0138**	**0.3383**	**0.0598**	**0.3031**	**0.0149**	**0.3472**	**0.0643**

Table 1: **Multi-view consistency.** Narrow/wide angles are $(0, 10]/[30, 60]$ degrees. Method-specific and common-mask scores are reported separately. Shared coverage (%): narrow 94.6, wide 91.2. † denotes per-time processing. Appendix B.4 defines masks and coverage.

	Artistic quality		Nine pairs		Blind rating	
Method	ArtScore↑	ArtFID↓	ArtScore↑	CM Loss↓	Style↑	Content↑
StyleGaussian (2024)†	0.583	42.7	*n/a*	*n/a*	*n/a*	*n/a*
StylizedGS (2025)†	0.641	35.2	*n/a*	*n/a*	*n/a*	*n/a*
FantasyStyle (2026)†	0.689	29.8	*n/a*	*n/a*	*n/a*	*n/a*
4DStyleGaussian (2024/2026)	0.712	26.4	0.698	0.817	3.41 ± 0.62	3.68 ± 0.54
Style4D (2026)	0.738	23.1	0.724	0.767	3.72 ± 0.51	3.91 ± 0.47
Ours	**0.781**	**19.6**	**0.769**	**0.686**	$\mathbf{4.18 \pm 0.39}$	$\mathbf{4.25 \pm 0.36}$

Table 2: **Quality and reference matching.** The first block reports single-sequence ArtScore and image-set ArtFID; the second averages nine scene–style pairs. Blind ratings are mean ± participant SD (1–5), with 24 raters. Appendices B.5 and B.6 detail pairs, rating intervals, and ArtFID sets; n/a denotes unevaluated entries.



Figure 4: **Style propagation through LEGO motion.** Each row shows a reference, the source, and five successive BindStyle frames. Circular motifs and angular brushwork follow the articulated bucket, chassis, and base.

	Method-specific masks				Common masks			
	Short-range		Long-range		Short-range		Long-range	
Method	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓
Per-frame extensions								
StyleGaussian (2024)†	0.2544	0.0356	0.4226	0.0787	0.2618	0.0379	0.4342	0.0835
StylizedGS (2025)†	0.2316	0.0298	0.3881	0.0652	0.2387	0.0318	0.3994	0.0694
FantasyStyle (2026)†	0.2090	0.0252	0.3530	0.0565	0.2161	0.0270	0.3638	0.0603
Native dynamic methods								
4DStyleGaussian (2024/2026)	0.1878	0.0214	0.3137	0.0485	0.1945	0.0229	0.3241	0.0518
Style4D (2026)	0.1702	0.0189	0.2898	0.0421	0.1768	0.0203	0.2995	0.0452
Ours	**0.1431**	**0.0147**	**0.2521**	**0.0347**	**0.1492**	**0.0158**	**0.2613**	**0.0374**

Table 3: **Temporal consistency.** Gaps are 2 and 20 frames. The right block uses one common mask per pair. Shared coverage (%): short 95.1, long 92.4. † denotes per-time processing. Appendix B.4 specifies the protocol.



Figure 5: **Geometry supervision and teacher conditions.** After reference and source, columns show VGG on SH, VGG on SH and geometry, joint removal of MotionAdapter and ControlNet, and BindStyle. Rows show Horse, Beagle, and Girlwalk.

4.3 Geometry Supervision and Gradient Routing

The ablations examine parameter capacity and supervisory assignment through distinct controls. Under VGG supervision, enabling geometry improves on SH-only editing in Table 4, supporting the value of editable Gaussian support. With both teachers present, SH-only editing also trails BindStyle, although this comparison changes both parameter scope and gradient routing. Fig. 5 illustrates the supervisory variants. These results concern rendered quality and consistency, not the magnitude or location of style-induced geometric changes. Appendix B.6 and Fig. 6 clarify the scope of the geometry evidence.

Removing MotionAdapter and ControlNet together tests their combined contribution. The unrouted control retains both teachers, initialization, and budget, but sends both losses to both parameter groups. Its higher aggregate errors in Table 4 support the supervisory assignment at the mean level rather than attributing the gain solely to parameter capacity. Appendix B.6 summarizes these differences without inferring case-wise improvements or statistical significance from aggregate statistics.

	Single case			Three matched cases		
Variant	CM↓	T-LPIPS↓	RMSE↓	CM↓	T-LPIPS↓	RMSE↓
VGG-only (SH-only)*	0.9743	0.2684	0.0391	0.981 ± 0.012	0.274 ± 0.008	0.041 ± 0.002
VGG + video (SH-only) *	0.9126	0.2415	0.0342	0.921 ± 0.014	0.248 ± 0.007	0.036 ± 0.002
VGG on SH and geometry w/o MotionAdapter and ControlNet	0.8567	0.2013	0.0276	0.869	0.208	0.029
VGG + video on SH and geometry*	0.9215	0.2358	0.0334	0.934	0.241	0.035
Ours*	0.7934	0.1782	0.0221	0.801 ± 0.015	0.183 ± 0.006	0.023 ± 0.001
	0.7082	**0.1426**	**0.0147**	$\mathbf{0.715 \pm 0.009}$	$\mathbf{0.146 \pm 0.004}$	$\mathbf{0.015 \pm 0.001}$

Table 4: **Geometry supervision and routing.** The same-teacher SH-only control retains both teachers but freezes geometry. In the three-case block, $*$ denotes three seeds per case; SD is across seed-wise case means. VGG + video on both groups is the unrouted control. Table 12 summarizes differences between the displayed aggregate means.

Tables 5–6 report separate style/seed ranges for each view and time group. Wider view baselines and longer temporal gaps shift these ranges upward, while style variation generally exceeds seed variation. These ranges describe successful runs, not a paired significance test; the inventory also records one failed seed run.

5 Limitations

The method requires an existing reconstruction and corresponding meshes at every requested time; it neither predicts motion nor edits mesh topology. Fixed face attachments do not prevent excessive local offsets, scales, or coverage, and the formulation contains no explicit surface regularizer. The geometry teacher operates on RGB, so its guidance can reflect appearance mismatch or reference content rather than desirable shape change. Depth conditioning may limit stylistic freedom, while short temporal windows and changing cameras restrict the temporal prior's applicability. Finally, SH view dependence and visibility changes can produce temporal variation despite shared parameters. These limitations necessitate separate evaluation of appearance, support geometry, and temporal behavior.

6 Conclusion

BindStyle learns shared Gaussian appearance and local geometry from one style image and transports them through known mesh motion. Parameter-routed optimization assigns VGG supervision to SH and video diffusion guidance to geometry. Multi-style comparisons favor its reference matching, while matched-teacher ablations support separating the supervisory roles. Inference uses the edited model without visual teachers or deformation prediction.

AI USE STATEMENT

Generative AI tools assisted with method and experiment discussions, code implementation and debugging, literature search, manuscript editing, and preparation of conceptual illustrations. Experimental measurements were supplied by the authors; AI tools also supported arithmetic and cross-reference checks. The authors remain responsible for the methods, measurements, and final content.

REFERENCES

Mahmoud Afifi, Marcus A. Brubaker, and Michael S. Brown. HistoGAN: Controlling colors of GAN-generated and real images via color histograms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7941–7950, 2021. URL https://arxiv.org/abs/2011.11731.

Ang Cao and Justin Johnson. HexPlane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 130–141, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cao_HexPlane_A_Fast_Representation_for_Dynamic_Scenes_CVPR_2023_paper.html.

Beiqi Chen, Shuai Shao, Haitang Feng, Jianhuang Lai, Jianlou Si, and Guangcong Wang. Style4D-Bench: A benchmark suite for 4d stylization. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 2769–2777, 2026. doi: 10.1609/aaai.v40i4.37266. URL https://ojs.aaai.org/index.php/AAAI/article/view/37266.

Junyu Chen, Jie An, Hanjia Lyu, Christopher Kanan, and Jiebo Luo. Learning to evaluate the artness of AI-generated images. *IEEE Transactions on Multimedia*, 26:10731–10740, 2024. URL https://arxiv.org/abs/2305.04923.

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*, 2026. URL https://arxiv.org/abs/2604.07795.

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.

Bruno Galerne, Jianling Wang, Lara Raad, and Jean-Michel Morel. SGSST: Scaling gaussian splatting style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26535–26544, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Galerne_SGSST_Scaling_Gaussian_Splatting_Style_Transfer_CVPR_2025_paper.html.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html.

Antoine Guédon and Vincent Lepetit. SuGaR: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.12775.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.04725.

Agrim Gupta, Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Characterizing and improving stability in neural style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4067–4076, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Gupta_Characterizing_and_Improving_ICCV_2017_paper.html.

Guilherme G. Haetinger, Jingwei Tang, Raphael Ortiz, Paul Kanyuk, and Vinicius Azevedo. Controllable neural style transfer for dynamic meshes. In *ACM SIGGRAPH Conference Papers*, 2024. URL https://studios.disneyresearch.com/2024/07/28/controllable-neural-style-transfer-for-dynamic-meshes/.

Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://instruct-nerf2nerf.github.io/.

Ming He, Zhixiang Chen, and Steve Maddock. SIC3D: Style image conditioned text-to-3d gaussian splatting generation. *arXiv preprint arXiv:2604.08760*, 2026. URL https://arxiv.org/abs/2604.08760.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2304.07090.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Real-Time_Neural_Style_CVPR_2017_paper.html.

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*, pp. 18342–18352, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Huang_StylizedNeRF_Consistent_3D_Scene_Stylization_As_Stylized_NeRF_via_2D-3D_CVPR_2022_paper.html.

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.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592433. URL https://arxiv.org/abs/2308.04079.

Áron Samuel Kovács, Pedro Hermosilla, and Renata Georgia Raidou. G-Style: Stylized gaussian splatting. *Computer Graphics Forum*, 43(7):e15259, 2024. doi: 10.1111/cgf.15259. URL https://diglib.eg.org/items/758761fc-b693-4242-bd3b-198bb4730ba5.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. URL https://arxiv.org/abs/2011.03277.

Wanlin Liang, Hongbin Xu, Weitao Chen, Feng Xiao, and Wenxiong Kang. 4DStyleGaussian: Generalizable 4d style transfer with gaussian splatting. *Pattern Recognition*, 172:112422, 2026. doi: 10.1016/j.patcog.2025.112422. URL https://www.sciencedirect.com/science/article/pii/S0031320325010830.

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. URL https://arxiv.org/abs/2509.26455.

Isabella Liu, Hao Su, and Xiaolong Wang. Dynamic gaussians mesh: Consistent mesh reconstruction from dynamic scenes. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LuGHbK8qTa.

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*, pp. 8338–8348, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Liu_StyleRF_Zero-Shot_3D_Style_Transfer_of_Neural_Radiance_Fields_CVPR_2023_paper.html.

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://kunhao-liu.github.io/StyleGaussian/.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2108.03647.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *International Conference on 3D Vision*, 2024. URL https://dynamic3dgaussians.github.io/.

Shaojie Ma, Yawei Luo, Wei Yang, and Yi Yang. MaGS: Reconstructing and simulating dynamic 3d objects with mesh-adsorbed gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2406.01593.

Yiqun Mei, Jiacong Xu, and Vishal M. Patel. ReGS: Reference-based controllable scene stylization with gaussian splatting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4035–4061, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/076c1fa639a7190e216e734f0a1b3e7b-Abstract-Conference.html.

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.

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. URL https://arxiv.org/abs/2307.01952.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14988.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2011.13961.

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

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*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *German Conference on Pattern Recognition*, pp. 26–36, 2016. URL https://arxiv.org/abs/1604.08610.

Abhishek Saroha, Florian Hofherr, Mariia Gladkova, Cecilia Curreli, Or Litany, and Daniel Cremers. ZDySS: Zero-shot dynamic scene stylization using gaussian splatting. *arXiv preprint arXiv:2501.03875*, 2025. URL https://arxiv.org/abs/2501.03875.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1409.1556.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations*, 2024. URL https://dreamgaussian.github.io/.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2003.12039.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1a87980b9853e84dfb295855b425c262-Abstract.html.

Matthias Wright and Björn Ommer. ArtFID: Quantitative evaluation of neural style transfer. In *German Conference on Pattern Recognition*, 2022. URL https://arxiv.org/abs/2207.12280.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2310.08528.

Yitong Yang, Xuexin Liu, Yinglin Wang, Jing Wang, Hao Dou, Changshuo Wang, and Shuting He. DiffStyle3D: Consistent 3d gaussian stylization via attention optimization. *arXiv preprint arXiv:2601.19717*, 2026a. URL https://arxiv.org/abs/2601.19717.

Yitong Yang, Yinglin Wang, Changshuo Wang, Huajie Wang, and Shuting He. FantasyStyle: Controllable stylized distillation for 3d gaussian splatting. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 11784–11792, 2026b. doi: 10.1609/aaai.v40i14.38164. URL https://ojs.aaai.org/index.php/AAAI/article/view/38164.

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. URL https://arxiv.org/abs/2308.06721.

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. StyleMaster: Stylize your video with artistic generation and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.07744.

Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.19415.

Dingxi Zhang, Yu-Jie Yuan, Zhuoxun Chen, Fang-Lue Zhang, Zhenliang He, Shiguang Shan, and Lin Gao. StylizedGS: Controllable stylization for 3d gaussian splatting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3604010. URL https://kristen-z.github.io/stylizedgs/.

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.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2302.05543.

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.