FIELDMORPH3D FOR TRAINING-FREE STRUCTURE-COHERENT 3D MORPHING VIA ATTENTION-CORRECTED PROGRESS FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D morphing generates continuous 3D transformations from a source object to a target object. However, large category or structural differences often produce duplicated parts, orientation inconsistency, and unstable local structures in intermediate states. Existing methods mainly focus on the interpolation or fusion of source and target features. They pay limited attention to structural evidence reliability and structure evolution consistency in the Sparse Structure (SS) stage. The SS stage determines spatial occupancy and coarse geometry. Errors at this stage can directly propagate to subsequent 3D decoding. Based on this observation, we propose FieldMorph3D, a training-free framework that corrects structural evidence before SS fusion and regulates the resulting deformation. A simplified attention model links shared query evidence to the Gram matrix of attention rows. Cross-Attention Feature Correction Gating (CFCG) uses a rank-one row update that provably reduces attention overlap for selected queries competing for the same key, followed by uncertainty gating. Latent Correction Fusion (LCF) stabilizes yaw across consecutive states. Occupancy and causal history first set voxel-wise fusion progress; the resulting provisional feature then exposes local structural inconsistency. LCF derives a displacement-penalized correction and selects its step size to avoid increasing fixed-graph energy while preserving source-to-target endpoint displacement. FieldMorph3D requires neither additional training nor explicit topology correspondence. It produces more coherent intermediate structures across challenging 3D morphing pairs.

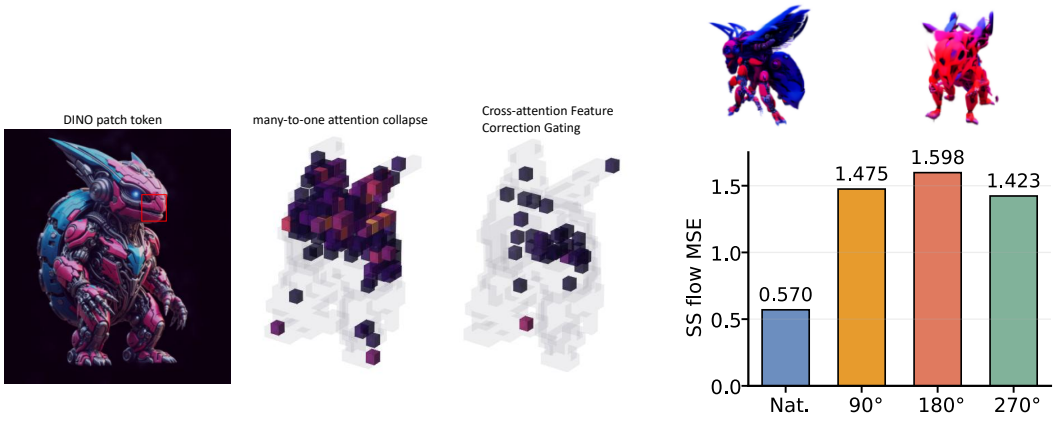
1 INTRODUCTION

Morphing generates a coherent transformation from a source object to a target object Wolberg (1996; 1998); Shechtman et al. (2010). It has been widely studied in both 2D image morphing Beier & Neely (1992); Lee et al. (1995); Zhang et al. (2024); Yang et al. (2024); Cao et al. (2025) and 3D shape morphing Kent et al. (1992); Alexa et al. (2000); Yang et al. (2025); Sun et al. (2026). Recent diffusion and flow-based generative models have greatly improved image morphing Ho et al. (2020); Song et al. (2021); Rombach et al. (2022); Liu et al. (2023b); Lipman et al. (2023). However, 3D morphing remains substantially more challenging, especially when the source and target have large category or structural gaps. Each intermediate state must preserve plausible geometry, spatial occupancy, pose, and structural continuity. Large structural differences can therefore produce duplicated parts, orientation inconsistency, disconnected structures, and implausible topology transitions Yang et al. (2025); Yin et al. (2025); Sun et al. (2026); Liu et al. (2026).

Recent generation-based 3D morphing methods can be roughly divided according to where the morphing path is constructed. The first category generates intermediate states in image or multi-view space Zhang et al. (2024); Yang et al. (2024); He et al. (2024); Cao et al. (2025); Huang et al. (2025), and then lifts them into 3D assets with image-to-3D models Liu et al. (2023a); Hong et al. (2024); Xu et al. (2024); Tochilkin et al. (2024); Tang et al. (2024); Gao et al. (2024). Independent 3D reconstruction may then cause geometry drift and temporal inconsistency.

The second category constructs the morphing path directly inside a 3D generative prior. Recent methods interpolate, transport, or fuse source and target states in 3D diffusion, flow, or structured

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.



(a) (b)

Figure 1: (a)Many SS queries respond strongly to one selected DINO patch. Repeated retrieval risks reusing its features across locations; CFCG reduces selected competing responses. (b)Basic fusion may favor a high-overlap yaw over the original orientation. In this example, SS flow loss favors natural yaw; correction before fusion avoids mixing incompatible features that a later rotation cannot disentangle.

latent spaces Yang et al. (2025); Yin et al. (2025); Sun et al. (2026); Liu et al. (2026). This route avoids repeated image-to-3D lifting, but its quality strongly depends on how intermediate structures are generated and regulated inside the generative model.

TRELLIS uses a structured representation for native 3D generation Xiang et al. (2025). Its Sparse Structure (SS) stage predicts active voxels that define the object's coarse spatial support. The subsequent Structured LATent (SLAT) stage predicts detailed features on this support and decodes the final 3D asset. Thus, SS largely determines the generated shape, and its structural errors are difficult to recover through later refinement. This makes SS critical for 3D morphing: every intermediate state needs a valid coarse structure before detailed geometry and appearance can be generated.

Recent TRELLIS-based morphing methods rely on attention to match and fuse source–target features Sun et al. (2026); Liu et al. (2026); Vaswani et al. (2017); Hertz et al. (2023); Cao et al. (2023). Attention can bring related source and target evidence to corresponding positions of a shared SS latent-token sequence without explicit topology correspondence. This makes fusion possible, but the retrieved evidence can be unreliable across large category gaps. Weak structural or semantic similarity can concentrate many spatial queries on a few dominant keys Martins & Astudillo (2016); Qiu et al. (2025); Darcet et al. (2024), transferring similar features to regions that should develop differently. The resulting SS features can still change yaw abruptly between states, conflict with their local neighborhoods, or evolve at an unsuitable common rate. Figure 1 illustrates attention concentration and yaw-dependent SS flow loss in two examples.

With TRELLIS weights frozen, we begin **FieldMorph3D** at the SS stage, where coarse structure shapes later geometry and appearance. We observe that cross-attention tends to gather related source and target features at corresponding positions in a shared sequence of latent tokens, providing an implicit basis for fusion without explicit geometric correspondence. Many queries can also concentrate on a few keys, an attention-sink-like pattern that sends ambiguous evidence to distinct regions. **Cross-Attention Feature Correction Gating (CFCG)** therefore corrects contested scores and gates uncertain responses before fusion. These two operations limit repeated retrieval from dominant keys and keep diffuse attention responses from becoming structural cues.

Once reliable evidence is co-located at corresponding token positions, fusion must turn that local matching into a coherent shape across states. A pose accepted by the frozen prior may break yaw continuity, and local mixing can disrupt coarse neighborhoods. **Latent Correction Fusion (LCF)** stabilizes orientation and protects the low-frequency SS scaffold while preserving the source-to-target displacement; SLAT later supplies fine geometry and appearance. Because structures emerge

and disappear at different rates, occupancy history also adjusts progress across SS tokens. FieldMorph3D thus connects score-level evidence correction with structure-aware fusion and sequence-level evolution within a frozen 3D prior.

The main contributions can be summarized as follows:

- We formulate two SS-stage bottlenecks through attention evidence overlap and spatial inconsistency, and derive the core corrections from low-rank matrix updates.
- We propose Cross-Attention Feature Correction Gating (CFCG), and show that its score update reduces overlap for selected queries competing for the same key.
- We derive LCF's rank-one local correction and an energy-controlled step that preserves endpoint displacement.

2 Related Work

2.1 Generative 3D Morphing

Recent generative 3D morphing methods reduce the dependence on explicit geometric correspondence by constructing deformation paths inside pretrained generative models Kilian et al. (2007); Ovsjanikov et al. (2012); Solomon et al. (2015); Sorkine & Alexa (2007); Sumner et al. (2007). Yang et al. (2025) introduced a regenerative morphing framework based on a 3D diffusion prior. The method interpolates multiple internal components of the diffusion process and improves cross-category transformations through feature fusion. MorphAny3D Sun et al. (2026) further exploits the structured latent space of TRELLIS. It uses attention-based source–target fusion, temporal feature reuse, and orientation correction to improve training-free 3D morphing. Interp3D Liu et al. (2026) introduces progressive alignment across condition, structure, and texture spaces. It improves source–target correspondence and fine-grained interpolation for textured 3D assets. These methods have significantly improved cross-category morphing by introducing stronger generative priors, structured latent representations, and correspondence-aware fusion. However, large structural gaps still make intermediate Sparse Structures difficult to stabilize. In particular, attention-based matching may provide repeated or unreliable structural evidence, while direct sparse-token fusion may introduce global misalignment, local structural inconsistency, and unstable structure evolution. Our SS-stage process corrects contested attention scores and gates uncertain evidence, then uses occupancy-dependent progress to form a provisional fusion whose local structure is corrected on the shared voxel grid.

2.2 Structured 3D Generative Priors

Large-scale pretrained models have greatly improved feed-forward 3D reconstruction and generation Choy et al. (2016); Fan et al. (2017); Groueix et al. (2018); Wang et al. (2018); Park et al. (2019); Mescheder et al. (2019); Mildenhall et al. (2020); Kerbl et al. (2023); Chan et al. (2022); Gao et al. (2022). LRM Hong et al. (2024) introduced a scalable transformer-based framework for single-image 3D reconstruction and demonstrated strong generalization across diverse objects. InstantMesh Xu et al. (2024) combines multi-view generation with a sparse-view reconstruction model and efficiently produces high-quality 3D meshes from a single image. TRELLIS Xiang et al. (2025) further introduces a structured 3D latent representation and separates generation into a Sparse Structure (SS) stage and a Structured LATent (SLAT) stage. The SS stage determines active spatial support and coarse object shape, while the SLAT stage represents detailed geometry and appearance. These generative priors mainly focus on producing a single final 3D asset. They do not directly regulate the continuous evolution between two structurally different endpoints. TRELLIS's SS–SLAT decomposition regulates matching, fusion, and evolution before SS errors reach decoding.

3 Method

Frozen TRELLIS supplies endpoint SS features $H_s, H_t \in \mathbb{R}^{N\times d}$ on a common voxel grid. Cross-attention retrieves endpoint evidence through $O = AV$, with query-to-key probabilities $A \in \mathbb{R}^{N\times K}$

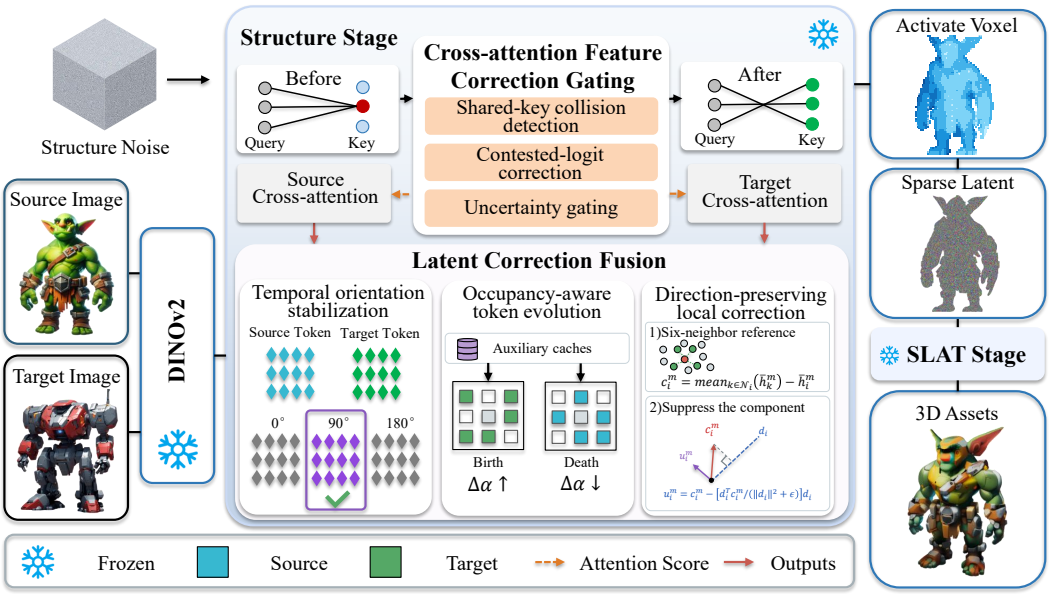


Figure 2: FieldMorph3D operates in the sparse-structure (SS) stage of frozen TRELLIS. CFCG reduces competition for cross-attention evidence; LCF stabilizes orientation, sets voxel-wise fusion progress, and corrects provisional features before SLAT decoding.

and values $V \in \mathbb{R}^{K\times d}$. CFCG corrects selected attention collisions. LCF stabilizes yaw, sets voxel-wise progress, and corrects the resulting provisional SS fusion before decoding fine geometry and appearance with frozen TRELLIS weights.

3.1 Cross-Attention Feature Correction Gating

For query i and key j, the attention score and retrieved feature are:

$$S_{ij} = \frac{q_i^\top k_j}{\sqrt{d}}, \qquad A_{i,:} = \mathrm{softmax}(S_{i,:}), \qquad o_i = \sum_j A_{ij} v_j. \tag{1}$$

To isolate the effect of attention, consider a simplified model in which value perturbations are conditionally independent across keys, have zero mean, and share covariance $\Sigma \succeq 0$. Then:

$$\mathrm{Cov}(o_i, o_k \mid A) = \sum_j A_{ij} A_{kj} \Sigma = (AA^\top)_{ik} \Sigma. \tag{2}$$

In this model, the off-diagonal entries of $AA^\top$ measure shared fluctuations in retrieved evidence. We therefore focus on queries that compete for the same dominant key.

Let $j_i = \arg\max_j S_{ij}$ and $s_i = S_{i,j_i}$. For each key, define its number of top-key assignments, strongest response, and weaker positive competitors:

$$c_j = |\{i : j_i = j\}|, \quad w_j = \max_{i:j_i=j} s_i, \quad \mathcal{L}_j = \{i : j_i = j,\ c_j > 1,\ 0 < s_i < w_j\}. \tag{3}$$

For $i \in \mathcal{L}_j$, CFCG corrects only the contested logit:

$$\widetilde{S}_{ij} = [S_{ij} - w_j]_+ = 0; \qquad \widetilde{S}_{ik} = S_{ik} \quad (k \neq j). \tag{4}$$

To see why lowering this logit is useful, let w be the winning query for key j and $q = A_{w,:}$ its attention row. Write the losing row as $a(p) = p\mathbf{e}_j + (1-p)r$, where $p = A_{ij}$, $\mathbf{e}_j$ is the key-space basis vector, and r is the conditional distribution over the remaining keys. Holding those keys' logits fixed gives:

$$\phi(p) = a(p)^\top q, \qquad \frac{d\phi}{dp} = q_j - r^\top q \geq 0, \tag{5}$$

because j is also the winner's top key. The softmax probability p increases with its logit, so setting that logit to its lower bound of zero minimizes the overlap over $[0, S_{ij}]$. For the selected queries, this is exactly the update in Eq. (4).

If $p' < p$ is the probability after correction, the change to the attention row and its Gram entry is:

$$\widetilde{a}_i - a_i = (p' - p)(\mathbf{e}_j - r), \qquad (\widetilde{A}\widetilde{A}^\top - AA^\top)_{iw} = (p' - p)(q_j - r^\top q) \le 0. \tag{6}$$

The update is rank one, since $\widetilde{A} - A = \mathbf{e}_i(\widetilde{a}_i - a_i)^\top$ for one selected query. Summing Eq. (6) over the selected winner–loser pairs shows that their total overlap cannot increase. Under Eq. (2), the corresponding conditional cross-covariance also decreases in positive-semidefinite order. The proof and scope of this pairwise result are given in Appendix A.

After score correction, we recompute $\widetilde{A} = \mathrm{softmax}(\widetilde{S})$ and gate uncertain responses:

$$e_i = -\sum_j \widetilde{A}_{ij} \log(\widetilde{A}_{ij} + \epsilon), \quad \gamma_i = \max_j \widetilde{A}_{ij}, \quad g_i = \mathbf{1}[e_i \le \tau_e \text{ or } \gamma_i \ge \tau_\gamma], \quad \widetilde{o}_i = g_i \sum_j \widetilde{A}_{ij} v_j. \tag{7}$$

The gate rejects diffuse, low-confidence evidence; the overlap result concerns the score correction. For each branch $b \in \{s, t\}$, the unchanged attention output projection and frozen block operations map $\widetilde{o}_{b,i}$ to $\widetilde{h}_{b,i}$. Their final voxel features form $\widetilde{H}_b$, the input to LCF.

3.2 Latent Correction Fusion

Temporal orientation stabilization addresses the dominant quarter-turn yaw ambiguity in TRELLIS SS outputs. For branch $b \in \{s, t\}$, let $p_b^{m,r}$ be its descriptor pooled from frozen DINO features at state m after candidate yaw rotation r. Compare it with the preceding state in its already selected orientation, $p_b^{m-1,*}$:

$$r_b^{m,*} = \arg \min_{r \in \{0, \pi/2, \pi, 3\pi/2\}} \|p_b^{m,r} - p_b^{m-1,*}\|_2^2. \tag{8}$$

The unrotated first state initializes this four-way search; the selected rotation is applied before fusion. Local progress uses cached endpoint occupancy $o_{s,i}, o_{t,i} \in \{0, 1\}$ and previous binary occupancy o_i^j. For $m \ge 1$, let nonempty history $\mathcal{H}_m \subseteq \{0, \dots, m-1\}$ contain $m-1$, with $m_0 = \min \mathcal{H}_m$ and $\bar{o}_i^m = |\mathcal{H}_m|^{-1} \sum_{j \in \mathcal{H}_m} o_i^j$. Birth, death, and endpoint-neighbor support are:

$$b_i = o_{t,i}(1 - o_{s,i}), \quad d_i^o = o_{s,i}(1 - o_{t,i}), \quad s_i = \frac{b_i \sum_{k \in \mathcal{N}_i^6} o_{t,k} + d_i^o \sum_{k \in \mathcal{N}_i^6} o_{s,k}}{6}. \tag{9}$$

Here $\mathcal{N}_i^6$ contains the six coordinate neighbors, with missing sites counted as zero. For prescribed global progress $\alpha_m \in [0, 1]$, let $[x]_a^b = \min\{b, \max\{a, x\}\}$. The voxel-wise coefficient is:

$$\begin{aligned} \beta_i^m = \Big[\alpha_m + \Big[s_i \big(b_i(1 - o_i^{m-1} - \bar{o}_i^m) + d_i^o(o_i^{m-1} + \bar{o}_i^m - 1) \\ + o_i^{m-1} - o_i^{m_0} \big) \Big]_{-1}^{1} \Big]_0^1. \end{aligned} \tag{10}$$

This occupancy rule is heuristic; it uses no learned parameters and does not ensure monotonic β_i^m. Persistent voxels have $b_i = d_i^o = 0$ and follow α_m. Once β_i^m is fixed, form the *provisional* fusion on the common voxel grid:

$$d_i = h_{t,i} - h_{s,i}, \qquad \bar{h}_i^m = (1 - \beta_i^m) h_{s,i} + \beta_i^m h_{t,i} = h_{s,i} + \beta_i^m d_i. \tag{11}$$

Inspect $\bar{H}^m = [(\bar{h}_i^m)^\top]_{i=1}^N$. For nonempty active six-neighbor sets $\mathcal{N}_i$, let $n_i^m = |\mathcal{N}_i|^{-1} \sum_{k \in \mathcal{N}_i} \bar{h}_k^m$ and $c_i^m = n_i^m - \bar{h}_i^m$; set $c_i^m = 0$ for isolated voxels. On this fixed undirected graph, with unnormalized Laplacian L and each edge counted once, define:

$$E_G(H) = \tfrac{1}{2} \operatorname{tr}(H^\top L H) = \tfrac{1}{2} \sum_{(i,k) \in \mathcal{E}} \|h_i - h_k\|_2^2. \tag{12}$$

Its gradient at the provisional fusion is $(L\bar{H}^m)_i = -|\mathcal{N}_i| c_i^m$. Penalizing motion along the endpoint displacement gives a closed-form local correction:

$$u_i^m = \arg\min_u \left\{ \tfrac{1}{2} \|u - c_i^m\|_2^2 + \tfrac{1}{2\epsilon} (d_i^\top u)^2 \right\} = \left(I - \frac{d_i d_i^\top}{\|d_i\|_2^2 + \epsilon} \right) c_i^m. \tag{13}$$

The matrix in parentheses leaves components perpendicular to d_i unchanged and suppresses the parallel component. Here $\epsilon > 0$ is a fixed stabilizing constant.

With fixed thresholds τ_c, τ_n, activate correction for fusion cancellation or neighborhood deviation:

$$\chi_i^m = \mathbf{1}\left[\frac{\|\bar{h}_i^m\|_2}{(1-\beta_i^m)\|h_{s,i}\|_2 + \beta_i^m\|h_{t,i}\|_2 + \epsilon} < \tau_c \text{ or } \|c_i^m\|_2 > \tau_n\right]. \tag{14}$$

The clipped update is:

$$z_i^m = \chi_i^m u_i^m \min\left(1, \frac{\|\bar{h}_i^m\|_2 + \epsilon}{\|u_i^m\|_2 + \epsilon}\right). \tag{15}$$

The projected update $Z^m = [(z_i^m)^\top]_i$ is a non-increasing first-order direction for E_G at $\bar{H}^m$ (Appendix A). Minimize the scalar quadratic, choosing $\eta_m = 0$ if it is constant:

$$\eta_m \in \arg\min_{0 \le t \le 1} E_G(\bar{H}^m + tZ^m). \tag{16}$$

This computed step guarantees $E_G(\bar{H}^m + \eta_m Z^m) \le E_G(\bar{H}^m)$ on the fixed graph (Appendix A). We add the same $\eta_m z_i^m$ to both pre-fusion endpoint features. Their difference remains d_i, and final fusion gives:

$$h_i^m = \bar{h}_i^m + \eta_m z_i^m. \tag{17}$$

Thus β_i^m controls progress along d_i, while $\eta_m z_i^m$ corrects the diagnosed local feature.

4 Experiments

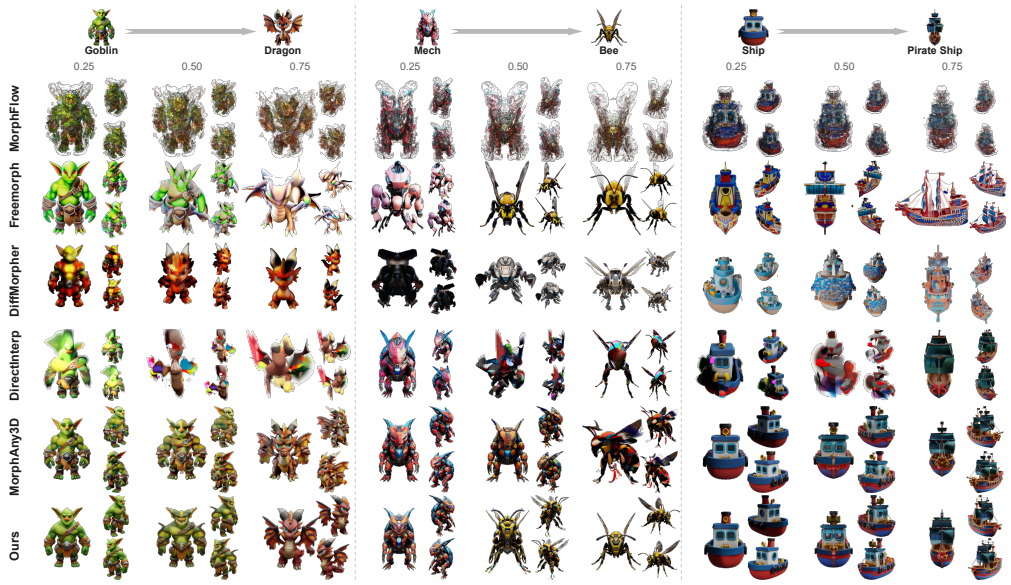


Figure 3: Qualitative comparison across three representative morphing settings. FieldMorph3D produces more coherent intermediate structures across different semantic, structural, and appearance variations.

4.1 Experimental Setup

Implementation Details. We evaluate FieldMorph3D on challenging cross-category source–target pairs using the image-to-3D version of TRELLIS Xiang et al. (2025). The 3D generator remains

frozen throughout all experiments, and no additional training or pair-specific optimization is required. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Unless otherwise specified, each morphing sequence contains 50 frames.

Baselines. We select representative baselines covering the major paradigms of recent generative morphing methods. DiffMorpher Zhang et al. (2024) and FreeMorph Cao et al. (2025) represent strong diffusion-based 2D morphing methods whose intermediate images are lifted to 3D. DirectInterp provides a direct structured-latent interpolation baseline and evaluates the behavior of the frozen 3D prior under simple source–target fusion. MorphFlow Tsai et al. (2022) represents flow-based multiview morphing, while MorphAny3D Sun et al. (2026) represents recent training-free morphing based on the structured latent space of TRELLIS. These baselines cover image-space generation, direct latent interpolation, native flow-based morphing, and structured-latent fusion. Additional comparisons with matching-based and latent-interpolation methods are provided in the supplementary material.

Evaluation Metrics. We evaluate intermediate plausibility and transition consistency using FID, PPL, PDV, and AS. **FID** measures the distributional realism of intermediate 3D assets. For each source–target pair, we render the source, target, and every intermediate asset from 12 uniformly sampled yaw views. InceptionV3 features are extracted from all rendered images, and FID is computed between the endpoint-reference distribution and the intermediate-result distribution. Lower FID indicates more plausible intermediate assets. **PPL** measures perceptual change along the morphing path. We render each sequence from six uniformly sampled yaw views and uniformly select 17 states from the full sequence. For each view, LPIPS is computed between 16 adjacent frame pairs and summed along the path. The final PPL is averaged over all views and source–target pairs. Lower PPL indicates smoother perceptual transitions. **PDV** measures the temporal uniformity of the morphing path. Using the same 16 adjacent LPIPS distances, we compute their variance independently for each view and then average across views and source–target pairs. Lower PDV indicates more uniform changes across the sequence. **AS** denotes the Aesthetic Selection Rate obtained with a vision-language evaluator. For each pair, AS records the highest-scoring method and reports its selection percentage over all 50 pairs.

4.2 Evaluation

Qualitative evaluation. Figure 3 presents three representative source–target pairs selected according to category gap and spatial occupancy gap. The first column contains objects with a large category gap but similar spatial occupancy, the second contains large gaps in both category and spatial occupancy, and the third contains objects from the same category with similar spatial occupancy but different styles. These settings cover semantic transformation, structural reorganization, and style variation. FieldMorph3D maintains clearer and more coherent intermediate structures across all three settings, especially under large semantic and spatial changes.

Quantitative evaluation. Table 1 reports results on 50 source–target morphing pairs covering more than 70 endpoint categories. The assets include samples from the TRELLIS evaluation data and a small number of additional assets reconstructed from GPT-generated RGBA images. All methods are evaluated using the same source–target pairs, rendering settings, and metrics. FieldMorph3D achieves stronger overall performance in intermediate plausibility and transition consistency, confirming the effectiveness of CFCG and LCF for SS-stage morphing.

Table 1: Quantitative comparison. Best results are bold.

Method	FID↓	PPL↓	PDV↓	AS (%)↑
DiffMorpher (2024)	208.08	6.65	0.0021	4.00
FreeMorph (2025)	164.68	5.89	0.0027	8.00
DirectInterp	150.94	3.72	0.0039	0.00
MorphFlow (2022)	284.96	2.41	0.0009	0.00
MorphAny3D (2026)	111.95	2.47	0.0006	38.00
Ours	**97.03**	**1.96**	**0.0005**	**50.00**



Figure 4: Module ablations. Left: CFCG attention correction. Right: LCF local correction and progress regulation. Each panel compares intermediate states with and without the corresponding module.

4.3 Ablation Study

Figure 4 compares the two modules. Without CFCG, competing queries can retrieve repeated structural evidence, causing unstable intermediate parts in the cross-category example. Without LCF, the same global interpolation rate is applied to persistent, emerging, and disappearing regions, leaving residual source structure and premature target parts. The complete method reduces both errors.

Table 2 reports the module-level results. CFCG changes the conditional evidence retrieved by SS queries, whereas LCF acts on the resulting spatial features and their progression. Relative to the Base variant, CFCG lowers FID from 197.72 to 106.40 and PDV from 0.0007 to 0.0005, while PPL rises from 2.47 to 2.74. LCF instead reduces PPL to 1.54 while lowering FID to 119.06. Their combination yields the best FID (97.03) and a tied-best PDV (0.0005), with PPL at 1.96. This pattern suggests that structural plausibility and perceptual path length place different demands on the intermediate SS states.

Table 2: Ablation study on key components of FieldMorph3D.

Variant	FID↓	PPL↓	PDV↓
Base	197.72	2.47	0.0007
CFCG	106.40	2.74	0.0005
LCF	119.06	**1.54**	0.0006
CFCG + LCF	**97.03**	1.96	**0.0005**

4.4 Applications of 3D Morphing

The same fusion machinery can also be driven by endpoint conditions that are not strict morphing targets. Figure 5 replaces the target condition with different style references while keeping the same source head. Because CFCG only regulates how endpoint evidence is retrieved and LCF only regulates how the fused structure evolves, the source head still supplies the structural scaffold of every intermediate state, while the style reference steers its geometric and material appearance. A recognizable head structure is retained throughout the sequence, and each style reference produces a distinct appearance path. This indicates that the SS-stage corrections extend beyond paired morphing to appearance-guided generation; a systematic evaluation against editing-specific baselines is left to future work.



Figure 5: Style-guided morphing from the same source head. Different style references produce distinct geometric and material appearances, while a recognizable head structure is retained throughout the intermediate sequence.

5 Conclusion

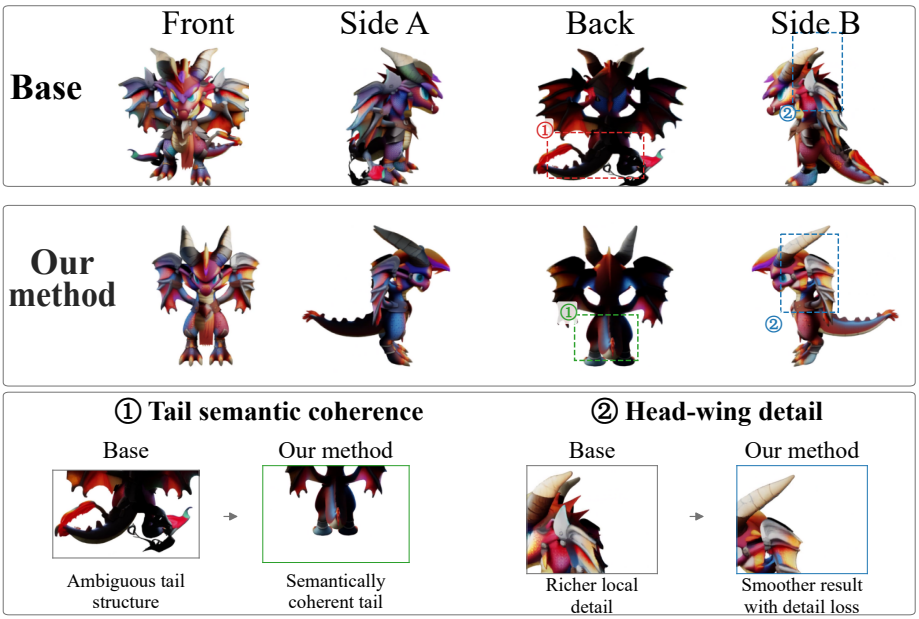


Figure 6: Limitation of stronger structure correction. Additional refinement makes the tail more coherent (1), but smooths some head and wing detail (2).

FieldMorph3D improves 3D morphing in the sparse-structure stage of a frozen generative prior. CFCG reduces competing attention evidence, while LCF stabilizes yaw across states and corrects local features without changing their source-to-target displacement. Occupancy-aware progress further adjusts where structures appear and disappear. The experiments show improved intermediate plausibility and temporal consistency without training or explicit topology correspondence.

The method still trades local detail for structural consistency. Figure 6 shows that additional refinement resolves an ambiguous tail but smooths the head and wing. Better separation of structural errors from meaningful high-frequency detail is an important direction for future work. Selective refinement of only the inconsistent regions may help, but remains to be evaluated. The yaw search also uses four discrete candidates, and the correction thresholds are fixed across pairs. Finer pose differences and threshold sensitivity across asset categories require further evaluation. Our experiments use image-conditioned mesh generation; text-conditioned morphing and other TRELLIS output representations remain to be tested.

REPRODUCIBILITY STATEMENT

The method and frozen backbone are specified in Section 3 and Appendix B. Appendix B.3 lists the shared hyperparameters, while Appendix C describes the evaluation protocol and metrics. All reported comparisons use the same source–target pairs and rendering settings.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting, editing, and adaptation to the ICLR 2027 LaTeX format. The authors take responsibility for the final text, references, figures, and scientific claims.

REFERENCES

Marc Alexa, Daniel Cohen-Or, and David Levin. As-rigid-as-possible shape interpolation. In *Proceedings of SIGGRAPH*, pp. 157–164, 2000.

Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *Proceedings of SIGGRAPH*, pp. 35–42, 1992.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22503–22513, 2023.

Yukang Cao, Chenyang Si, Jinghao Wang, and Ziwei Liu. FreeMorph: Tuning-free generalized image morphing with diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16123–16133, 2022.

Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pp. 628–644, 2016.

Timothee Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 605–613, 2017.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. GET3D: A generative model of high quality 3d textured shapes learned from images. In *Advances in Neural Information Processing Systems*, 2022.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. A papier-mache approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 216–224, 2018.

Qiyuan He, Jinghao Wang, Ziwei Liu, and Angela Yao. AID: Attention interpolation of text-to-image diffusion. In *Advances in Neural Information Processing Systems*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

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.

Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. MV-Adapter: Multi-view consistent image generation made easy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

James R. Kent, Wayne E. Carlson, and Richard E. Parent. Shape transformation for polyhedral objects. In *Proceedings of SIGGRAPH*, pp. 47–54, 1992.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023.

Martin Kilian, Niloy J. Mitra, and Helmut Pottmann. Geometric modeling in shape space. *ACM Transactions on Graphics*, 26(3):64, 2007.

Seung-Yong Lee, Kyung-Yong Chwa, Sung Yong Shin, and George Wolberg. Image metamorphosis using snakes and free-form deformations. In *Proceedings of SIGGRAPH*, pp. 439–448, 1995.

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

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023a.

Xiaolu Liu, Yicong Li, Qiyuan He, Jiayin Zhu, Wei Ji, Angela Yao, and Jianke Zhu. Interp3D: Correspondence-aware interpolation for generative textured 3d morphing. *arXiv preprint arXiv:2601.14103*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023b.

Andre F. T. Martins and Ramon Fernandez Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pp. 1614–1623, 2016.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.

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.

Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: A flexible representation of maps between shapes. *ACM Transactions on Graphics*, 2012.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 165–174, 2019.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2505.06708. URL https://arxiv.org/abs/2505.06708.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn 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.

Eli Shechtman, Alex Rav-Acha, Michal Irani, and Steven M. Seitz. Regenerative morphing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 615–622, 2010.

Justin Solomon, Fernando de Goes, Gabriel Peyre, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Eurographics Symposium on Geometry Processing*, pp. 109–116, 2007.

Robert W. Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics*, 26(3):80, 2007.

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.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, 2024.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. TripoSR: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

Chih-Jung Tsai, Cheng Sun, and Hwann-Tzong Chen. Multiview regenerative morphing with dual flows. In *Proceedings of the European Conference on Computer Vision*, pp. 492–509, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision*, pp. 55–71, 2018.

George Wolberg. Recent advances in image morphing. In *Proceedings of CG International*, 1996.

George Wolberg. Image morphing: A survey. *The Visual Computer*, 14(8–9):360–372, 1998.

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.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Songlin Yang, Yushi Lan, Honghua Chen, and Xingang Pan. Textured 3d regenerative morphing with 3d diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Zhaoyuan Yang, Zhengyang Yu, Zhiwei Xu, Jaskirat Singh, Jing Zhang, Dylan Campbell, Peter H. Tu, and Richard Hartley. IMPUS: Image morphing with perceptually-uniform sampling using diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Minghao Yin, Yukang Cao, and Kai Han. Wukong's 72 transformations: High-fidelity textured 3d morphing via flow models. In *Advances in Neural Information Processing Systems*, 2025.

Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7912–7921, June 2024.

A Derivations for the Low-Rank Corrections

Assumptions and attention covariance. Fix a cross-attention matrix A with at least two keys. To isolate the effect of its rows, assume that value perturbations satisfy $\mathbb{E}[v_j \mid A] = 0$ and $\mathrm{Cov}(v_j, v_k \mid A) = \mathbf{1}[j = k]\Sigma$ for a common $\Sigma \succeq 0$. These assumptions describe a simplified conditional model; they are not needed for the deterministic overlap result below. With $o_i = \sum_j A_{ij} v_j$, bilinearity of covariance gives:

$$\mathrm{Cov}(o_i, o_k \mid A) = \sum_{j,\ell} A_{ij} A_{k\ell} \,\mathrm{Cov}(v_j, v_\ell \mid A) = (AA^\top)_{ik}\Sigma. \tag{18}$$

Proposition 1: selected attention overlap is non-increasing. Let j be the common top key of a winning query w and a selected losing query $i \in \mathcal{L}_j$. Write $q = A_{w,:}$, $p = A_{ij}$, and $r_k = A_{ik}/(1-p)$ for $k \neq j$, with $r_j = 0$. If the logit of key j is changed from $S_{ij} > 0$ to any $x \in [0, S_{ij}]$, all other logits stay fixed. The new losing row is $a(p(x)) = p(x)\mathbf{e}_j + (1 - p(x))r$, where:

$$p(x) = \frac{e^x}{e^x + C}, \qquad C = \sum_{k \neq j} e^{S_{ik}} > 0. \tag{19}$$

Since j is a top key of w, $q_j \geq q_k$ for every k. Thus $q_j - r^\top q \geq 0$, giving:

$$\frac{d}{dx}\{a(p(x))^\top q\} = p(x)(1 - p(x))(q_j - r^\top q) \geq 0. \tag{20}$$

The minimum over $[0, S_{ij}]$ is attained at $x = 0$, exactly the CFCG score correction for $i \in \mathcal{L}_j$. If the winner has a unique top key, the derivative is strictly positive and this minimum is unique. For $p' = p(0) < p$, the change in the losing row is $\delta_i = (p' - p)(\mathbf{e}_j - r)$ and $A' = A + \mathbf{e}_i \delta_i^\top$, where $\mathbf{e}_i$ and $\mathbf{e}_j$ are query- and key-space basis vectors. Therefore $\mathrm{rank}(A' - A) \leq 1$, and the Gram-entry change is:

$$(A'A'^\top)_{iw} - (AA^\top)_{iw} = \delta_i^\top q = (p' - p)(q_j - r^\top q) \leq 0. \tag{21}$$

The winning row is unchanged. Summing Eq. (21) over all selected winner–loser pairs, while keeping the original pairs fixed, proves that their total overlap $\mathcal{C}(A) = \sum_j \sum_{i \in \mathcal{L}_j} A_{i,:}^\top A_{w(j),:}$ cannot increase. By Eq. (18), each corresponding conditional cross-covariance change, defined as corrected minus original, is a nonpositive scalar multiple of Σ and is negative semidefinite. Other entries of $AA^\top$ need not decrease.

Proposition 2: occupancy-conditioned correction decreases fixed-graph energy. Fix the endpoint occupancy, causal history, and undirected voxel graph for state m. Equation (10) then fixes β_i^m before feature correction. Let $E_G(H) = \frac{1}{2}\,\mathrm{tr}(H^\top L H)$. For a vertex i with nonempty neighborhood, $\nabla_{h_i} E_G(\bar{H}^m) = (L\bar{H}^m)_i = -|\mathcal{N}_i| c_i^m$, where $c_i^m = n_i^m - \bar{h}_i^m$. The strictly convex local objective in Eq. (13) has first-order condition $(I + d_i d_i^\top/\epsilon)u_i^m = c_i^m$. Sherman–Morrison gives:

$$\left(I + \frac{d_i d_i^\top}{\epsilon}\right)^{-1} = I - \frac{d_i d_i^\top}{\|d_i\|_2^2 + \epsilon} = P_i^\epsilon. \tag{22}$$

The symmetric matrix P_i^ϵ has eigenvalue 1 on directions orthogonal to d_i and eigenvalue $\epsilon/(\|d_i\|_2^2 + \epsilon)$ along d_i, so it is positive semidefinite. The mask and clipping in Eq. (15) give $z_i^m = \rho_i^m \chi_i^m P_i^\epsilon c_i^m$ for some $\rho_i^m \in [0, 1]$. Consequently:

$$g_m = \frac{d}{dt} E_G(\bar{H}^m + tZ^m)\bigg|_{t=0} = \sum_i \nabla_{h_i} E_G(\bar{H}^m)^\top z_i^m = -\sum_i |\mathcal{N}_i| \rho_i^m \chi_i^m (c_i^m)^\top P_i^\epsilon c_i^m \leq 0. \tag{23}$$

The first-order result alone does not cover a full unit step. Since $L \succeq 0$, $q_m = \mathrm{tr}((Z^m)^\top L Z^m) = \sum_{(i,k)\in\mathcal{E}} \|z_i^m - z_k^m\|_2^2 \geq 0$, and the exact quadratic expansion is:

$$E_G(\bar{H}^m + \eta Z^m) - E_G(\bar{H}^m) = \eta g_m + \tfrac{1}{2}\eta^2 q_m. \tag{24}$$

For $q_m > 0$, Eq. (16) has the explicit solution $\eta_m = [-g_m/q_m]_0^1$. Thus the change is nonpositive, and it is strictly negative when $g_m < 0$. If $q_m = 0$, then $LZ^m = 0$ and hence $g_m = \mathrm{tr}((\bar{H}^m)^\top L Z^m) = 0$; setting $\eta_m = 0$ leaves the feature unchanged. This guarantee concerns the fixed graph at one state. The shared update preserves endpoint displacement exactly: $(h_{t,i} + \eta_m z_i^m) - (h_{s,i} + \eta_m z_i^m) = d_i$. Isolated voxels have $c_i^m = z_i^m = 0$ and require no local correction.

B Additional Implementation Details

B.1 Integration with Frozen TRELLIS

FieldMorph3D uses TRELLIS Xiang et al. (2025) without updating its parameters. Endpoint features, sparse structures, and occupancy maps are generated once and cached. The SS sampler has 25 steps, each with eight cross-attention blocks; CFCG is applied in every such block, and LCF acts immediately before SS feature fusion. Each intermediate state runs an independent flow trajectory with the cached endpoint conditions. Only the occupancy history used by LCF passes between states; hidden flow features do not.

B.2 Sparse-Grid Coordinates and Endpoint Occupancy

SS flow features occupy a common fixed voxel lattice before final occupancy is decoded. Source and target features are therefore indexed by the same integer coordinates, without mesh correspondence or changes to TRELLIS voxelization. The cached endpoint occupancy maps may differ: a voxel active only at the target is a birth region, while one active only at the source is a death region. LCF uses these states to adjust local progress after CFCG has corrected the attention evidence.

B.3 Hyperparameters and Boundary Handling

All hyperparameters are fixed for every source–target pair. Table 3 lists the default configuration. Both local structural correction and occupancy-support estimation use the six-connected voxel neighborhood.

Table 3: Default hyperparameters used in all experiments.

Parameter	Meaning	Value
τ_e	Entropy threshold	6.0
τ_γ	Confidence threshold	1.0
τ_c	Collapse threshold	0.4
τ_n	Neighborhood threshold	0.4
ϵ	Projection regularizer	10^{-6}
$\mathcal{N}_i$	Voxel neighborhood	6-connected
M	Number of morphing states	50

The source endpoint serves as the first state, so no additional history initialization is required. Tokens without valid active neighbors are excluded from local correction. When the endpoint displacement $\|d_i\|_2$ is numerically zero, the direction-preserving correction is skipped at that location.

Temporal orientation stabilization compares each intermediate state with its preceding state using DINO-based descriptors. Equal rotation scores are rare except for strongly symmetric objects or large changes between adjacent states. We do not introduce a separate tie-breaking rule. Therefore, the initial endpoint orientation and highly ambiguous symmetries remain limitations of the current implementation.

C Evaluation Protocol and Metric Definitions

C.1 Definition of AS

We define AS as the *Aesthetic Selection Rate*. ChatGPT-5.4 is used as a vision-language evaluator to compare the perceptual and artistic quality of rendered 3D morphing videos. The evaluation considers overall harmony, surface detail and texture, artistic-style consistency, visual appeal, and inter-frame coherence.

Each evaluation group corresponds to one source–target pair and contains the results of all compared methods. Every morphing sequence contains 50 states, including the source and target endpoints. Each 3D state is rendered from six views. Selected intermediate states are additionally arranged

as overview views to support a holistic inspection of the full deformation process. The evaluator assigns a score to each method and selects the method with the highest score. The selected method receives one count for that evaluation group. We use the returned overall score directly and make no assumption about the evaluator's internal aggregation of views or frames.

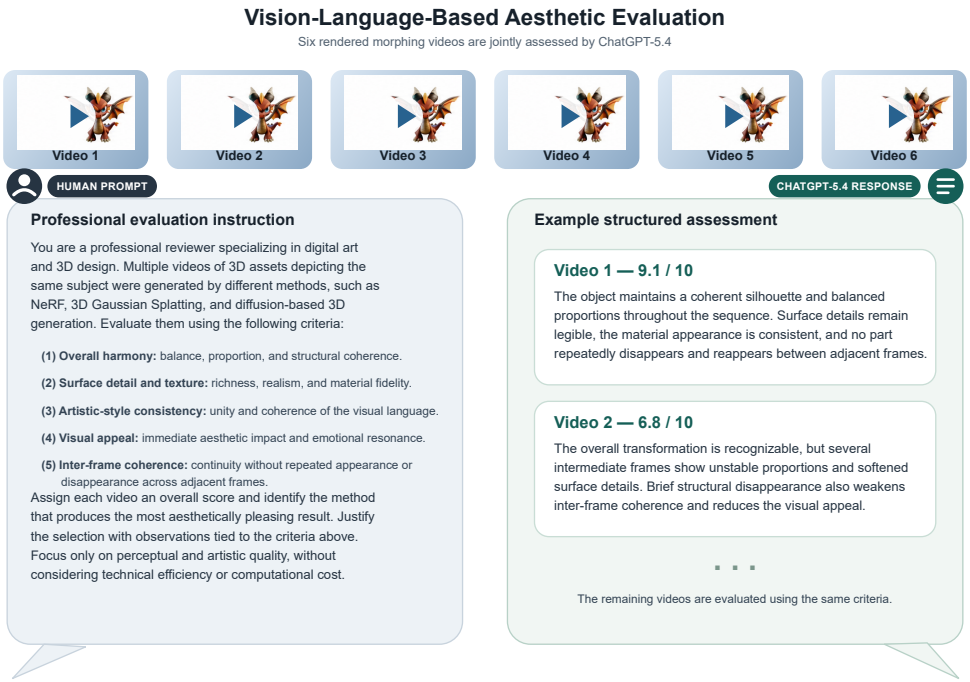


Figure 7: Illustration of the vision-language-based aesthetic evaluation. Six videos are presented with a fixed professional-review prompt. ChatGPT-5.4 assigns a score to each video and explains the assessment using the same five perceptual criteria. Only two example responses are shown for clarity.

For method k, AS is computed as:

$$\mathrm{AS}_k = \frac{1}{50} \sum_{q=1}^{50} \mathbb{I}\left[k = \arg\max_j s_{q,j}\right] \times 100\%,$$

where $s_{q,j}$ is the score assigned to method j for evaluation group q. Thus, AS measures how frequently a method is selected as the most aesthetically pleasing result rather than averaging its raw model scores. Figure 7 illustrates the evaluation interface and the structured form of the model response.

D Comparison with Interp3D

Figure 8 compares FieldMorph3D with Interp3D on three source–target pairs. For each pair, the same source and target endpoints are shown with five uniformly spaced intermediate states. FieldMorph3D maintains more coherent intermediate structures under changes in object category and spatial support. In particular, the biplane-to-helicopter and red-tree-to-staff examples show more gradual structural reorganization, while the castle-to-mushroom sequence retains a clearer central structure during the transition.

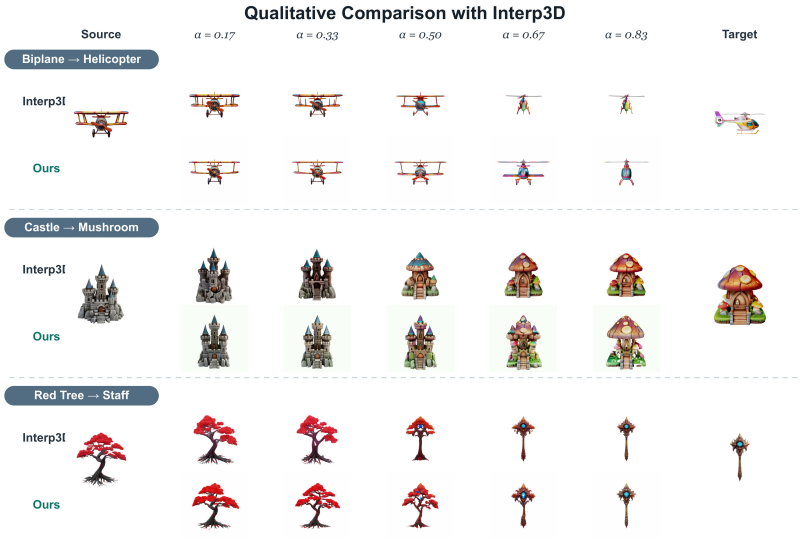


Figure 8: Qualitative comparison with Interp3D on three representative source–target pairs. Both methods are shown at the same five morphing progress values. Source and target endpoints are shared across the two rows of each example.

E Applications of 3D Morphing

FieldMorph3D can reuse different endpoint cues and can be applied to 3D assets beyond the primary evaluation setting. We present two examples covering dual-target control and morphing between complex 3D assets; style-guided morphing is shown in Figure 5 in the main text.

E.1 Dual-Target Morphing

Figure 9 separates structural and detailed target cues. The goblin provides the target structure, while the armored character provides appearance details. The resulting sequence gradually follows the target structure while incorporating the desired detail characteristics.

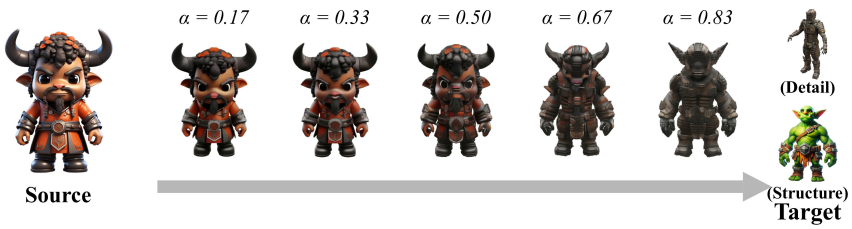


Figure 9: Dual-target morphing. The structure and detail conditions are provided by two different target assets.

E.2 Morphing with Alternative 3D Assets

Figure 10 further applies the morphing process to a complex robot-to-character pair. The intermediate states progressively reorganize the large mechanical structure into the target character while maintaining a continuous deformation path.

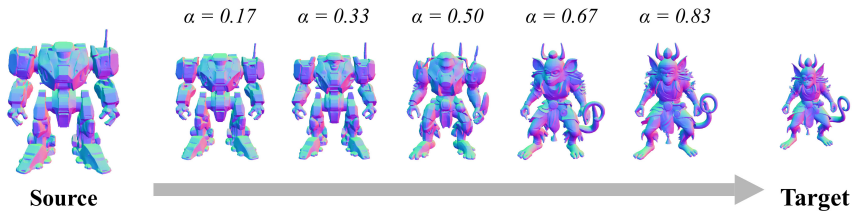


Figure 10: Morphing between complex 3D assets with large structural and semantic differences.

F Additional Generalization Results

Figure 11 presents a broader collection of morphing results across objects, characters, buildings, animals, plants, and vehicles. Each example includes the textured rendering and its corresponding normal rendering. The examples are collected from all result folders that contain both visualization modes, rather than being selected only from the main-paper cases. The consistent structural evolution across these diverse pairs provides additional evidence that FieldMorph3D generalizes beyond a small set of favorable source–target combinations.



Figure 11: Additional FieldMorph3D results on ten diverse source–target pairs. For every pair, the upper row shows textured renderings and the lower row shows the corresponding surface-normal renderings.