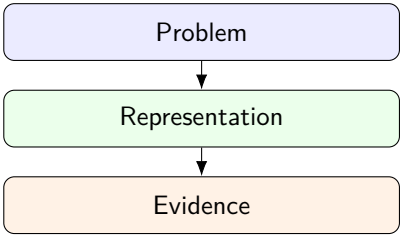
AniGen: Unified S^3 Fields for Animatable 3D Asset Generation

Yi-Hua Huang*, The University of Hong Kong, China
Zi-Xin Zou, VAST, China
Yuting He, The Chinese University of Hong Kong, China
Chirui Chang, The University of Hong Kong, China
Cheng-Feng Pu, Tsinghua University, China
Ziyi Yang, The University of Hong Kong, China
Yuan-Chen Guo, VAST, China
Yan-Pei Cao†, VAST, China
Xiaojuan Qi†, The University of Hong Kong, China

Talk Roadmap

- Generate animatable 3D assets from a single image, not just static meshes.
- Output geometry, skeleton, and skinning weights in one asset.
- Core idea: jointly generate shape, skeleton, and skin in one unified spatial representation.

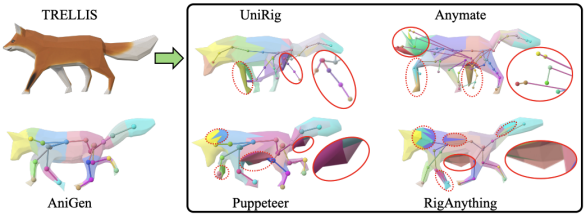


Static 3D Is Not Enough

- Games, VR/AR, embodied AI, and digital twins all need functional 3D assets.
- An animatable asset usually needs mesh geometry, a skeleton, and skinning weights.
- A static mesh is still closer to a statue than an object that can enter an animation workflow.

Generate-Then-Rig Is Brittle

- Generated meshes can contain fused limbs, pose bias, and local topology artifacts.
- Auto-rigging depends on clean topology and reliable structural cues.
- Small geometry errors can amplify into missing bones, wrong connectivity, and broken deformation.



What Must Be Generated?

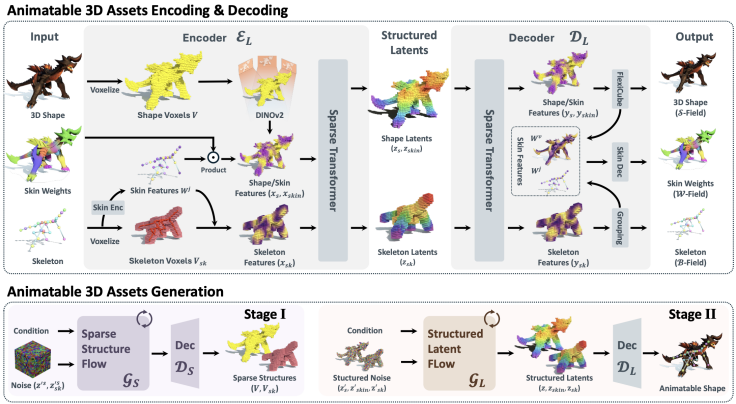
$$A = (M, B, W)$$

- M: mesh geometry
- B: skeleton joints and connectivity
- W: skinning weights

Why this is hard

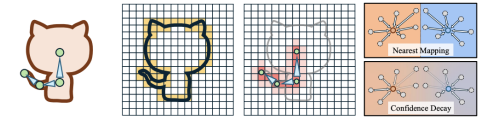
- Continuous geometry
- Irregular graph
- Variable-size weights

Unified S^3 Fields



Shape and Skeleton Fields

- Shape field lives on sparse surface voxels and stores FlexiCubes geometry / appearance parameters.
- Skeleton field lives around bones, not just the surface.
- Each point predicts nearest joint and parent vectors; voting and clustering recover the graph.



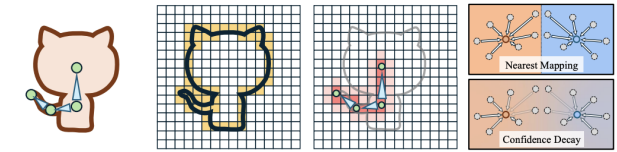
(a) Shape & Skeleton (b) Shape Voxels (c) Skeleton Voxels (d) Skeleton Field

Continuous to discrete

Field predictions → confidence voting → joint clustering

Confidence-Decaying Skeleton Field

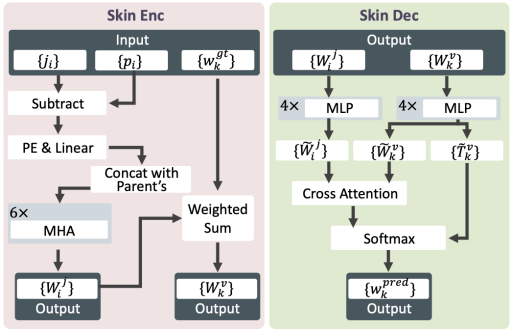
- Near Voronoi boundaries, the nearest joint identity can switch abruptly.
- Direct regression produces noisy skeletons and redundant bones.
- Confidence-weighted training and clustering focus learning on reliable regions.



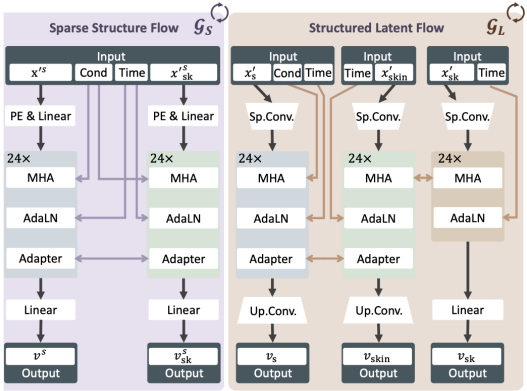
(a) Shape & Skeleton (b) Shape Voxels (c) Skeleton Voxels (d) Skeleton Field

Dual Skin Field + SkinAE

- Different assets have very different joint counts.
- Surface Skin Field describes vertex deformation features.
- Skeleton Skin Field describes joint influence features.
- SkinAE maps variable-cardinality skinning into a stable latent space.

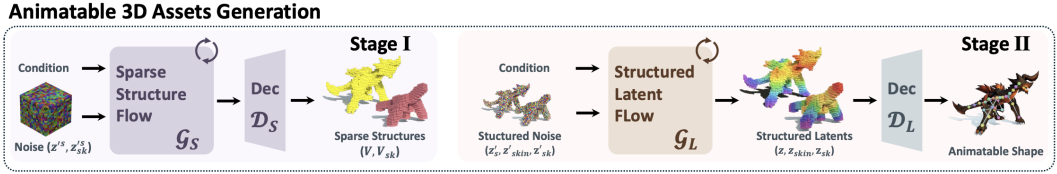


- Sparse Structure Auto-Encoder learns coarse shape/skeleton scaffold.
- Structured Latent Auto-Encoder encodes fine-grained fields.
- Stage I flow generates supports; Stage II flow generates structured latents.



From Image to Rigged Asset

- Image encoder provides conditioning features.
- Flow models generate support voxels and S^3 latents.
- Decoders recover mesh, joints/bones, and skinning weights.
- Output can be used by an animation pipeline.



Experimental Setup

- Dataset: ArticulationXL, about 33K rigged shapes from Objaverse-XL.
- Test set: 1K randomly sampled shapes.
- Baselines: TRELLIS + UniRig / Anymate / Puppeteer / RigAnything.
- Metrics cover skeleton structure, skinning quality, geometry, and runtime.

Main Quantitative Results

- Joint-GW measures structural similarity of skeleton graphs.
- Skin KL measures skinning distribution quality.
- The strongest gains are in rigging and deformation.

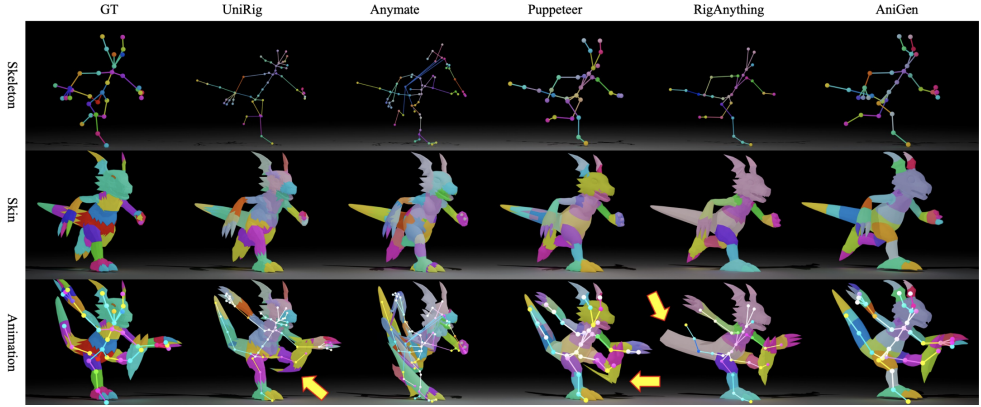
Method	Joint-GW ↓	Skin KL ↓
TRELLIS + UniRig	0.397	5.903
TRELLIS + Anymate	0.349	4.221
TRELLIS + Puppeteer	0.326	4.135
TRELLIS + RigAnything	0.383	6.451
AniGen	**0.286**	**2.919**

Geometry and Runtime Trade-Off

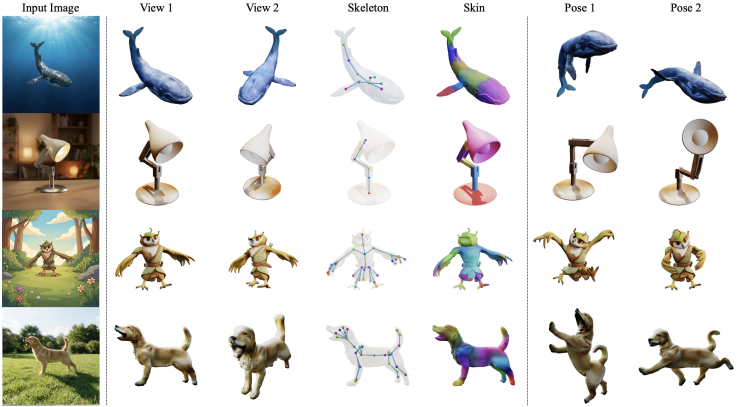
- AniGen is close to fine-tuned TRELLIS* on geometry metrics.
- Runtime is 19 s, close to the fastest sequential rigging baseline.
- The trade-off is small geometry loss for stronger rig consistency.

Method	Time
TRELLIS*	15 s
TRELLIS* + UniRig	146 s
TRELLIS* + Anymate	19 s
TRELLIS* + Puppeteer	36 s
TRELLIS* + RigAnything	127 s
AniGen	**19 s**

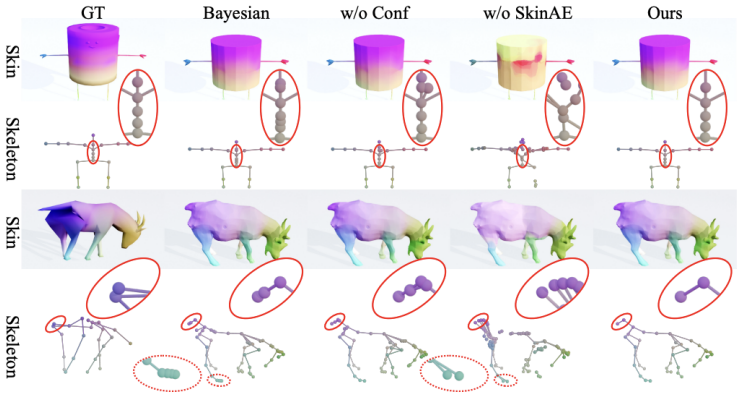
Qualitative Results: Rigging Comparisons



Qualitative Results: In-the-Wild Inputs



Qualitative Results: Ablation Examples



Ablation Study

- Removing confidence weakens skeleton quality.
- Bayesian uncertainty is weaker than explicit confidence supervision.
- Removing SkinAE hurts both skeleton and skinning metrics.

Method	Joint-GW ↓	Skin KL ↓
Bayesian confidence	0.310	3.174
w/o confidence	0.337	3.187
w/o SkinAE	0.383	5.138
Ours	**0.286**	**2.919**

Takeaways

1. S^3 fields give shape, skeleton, and skin a shared spatial representation.
2. Confidence-decaying skeleton field handles ambiguous bone boundaries.
3. Dual Skin Field + SkinAE handles variable joint counts.
4. AniGen treats animatability as a generation target, not as post-processing.

Limitations

Single-image conditioning; no explicit joint limits or physical properties yet.

Questions Remained

- If input becomes video instead of one image, how should S^3 fields change?
- Can the generated skeleton satisfy real DCC rigging workflows?
- Does embodied AI also need joint limits, physics, and collision proxies?
- Can this be combined with text-to-3D or multi-view reconstruction?

References

[1] Yi-Hua Huang, Zi-Xin Zou, Yuting He, Chirui Chang, Cheng-Feng Pu, Ziyi Yang, Yuan-Chen Guo, Yan-Pei Cao, and Xiaojuan Qi.
AniGen: Unified S^3 Fields for Animatable 3D Asset Generation.
arXiv:2604.08746v2, 2026.

[2] Project page: `https://yihua7.github.io/AniGen_web/`

[3] arXiv: `https://arxiv.org/abs/2604.08746`

[4] Code: `https://github.com/VAST-AI-Research/AniGen`