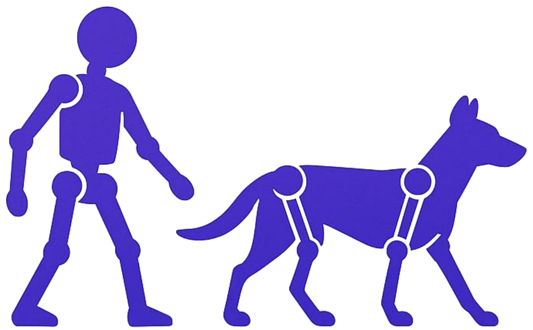
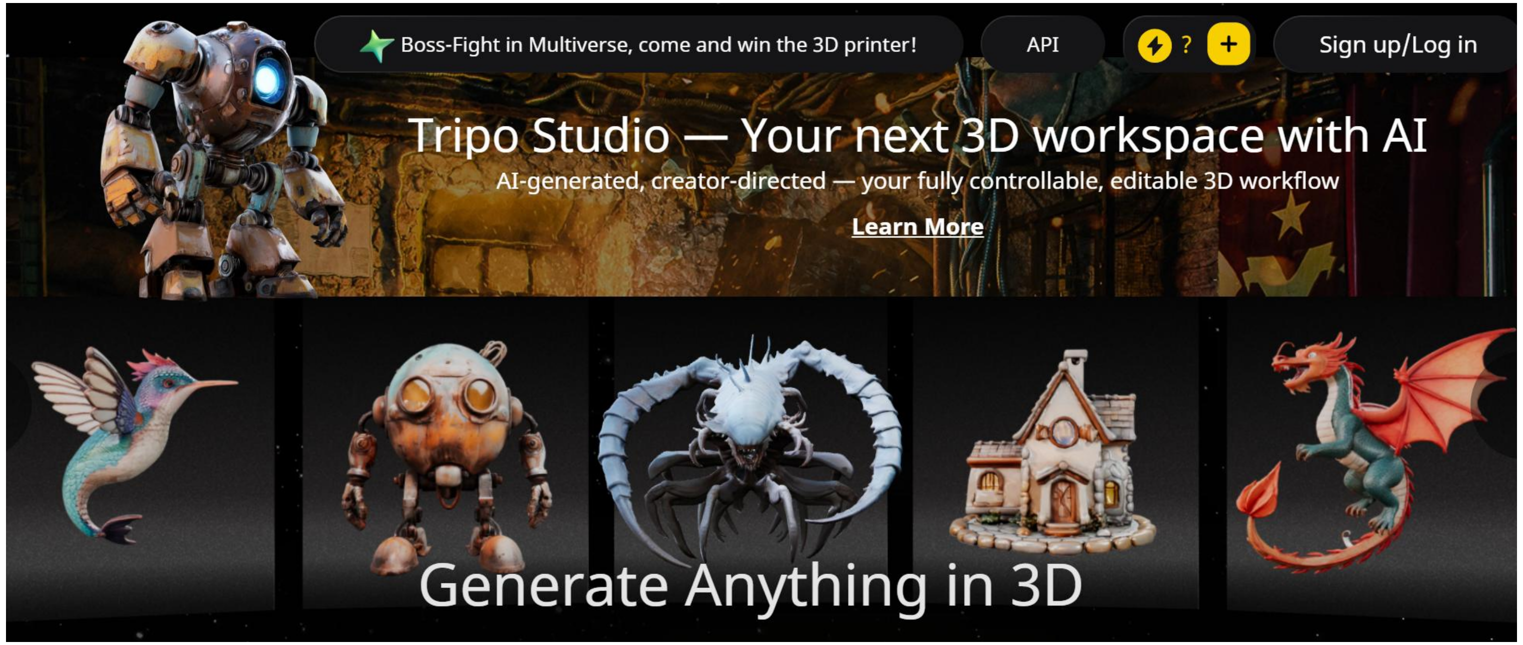


Rig and Animate Your 3D Models

Chaoyue Song
Nanyang Technological University

Why Rigging?

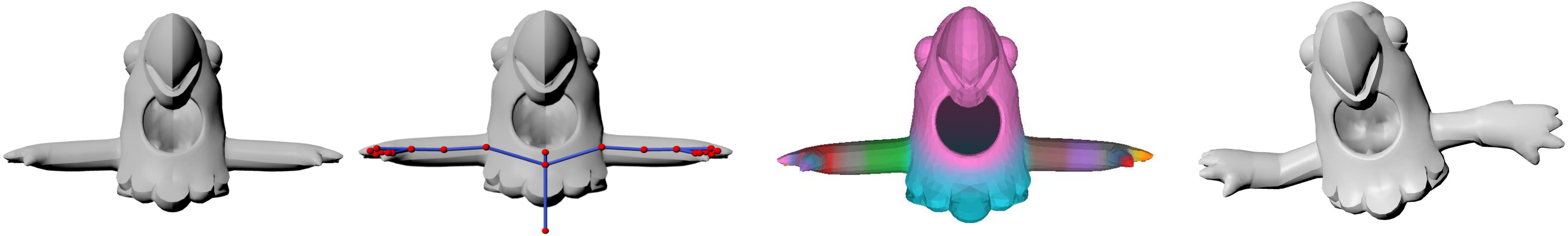
Clay



Tripo

Impressive geometry, texture, but... **Static**

Rigging definition



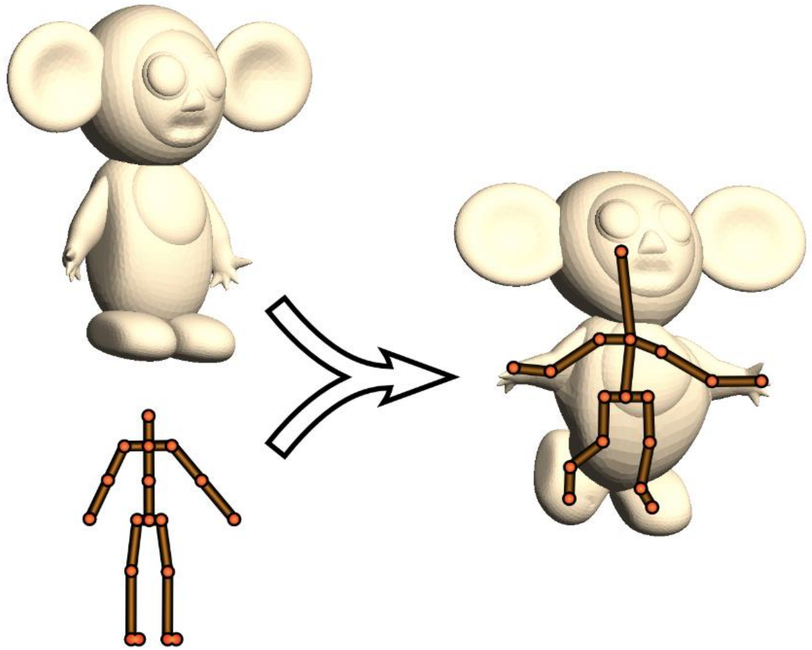
Input mesh | Skeleton | Skinning weights | Animation

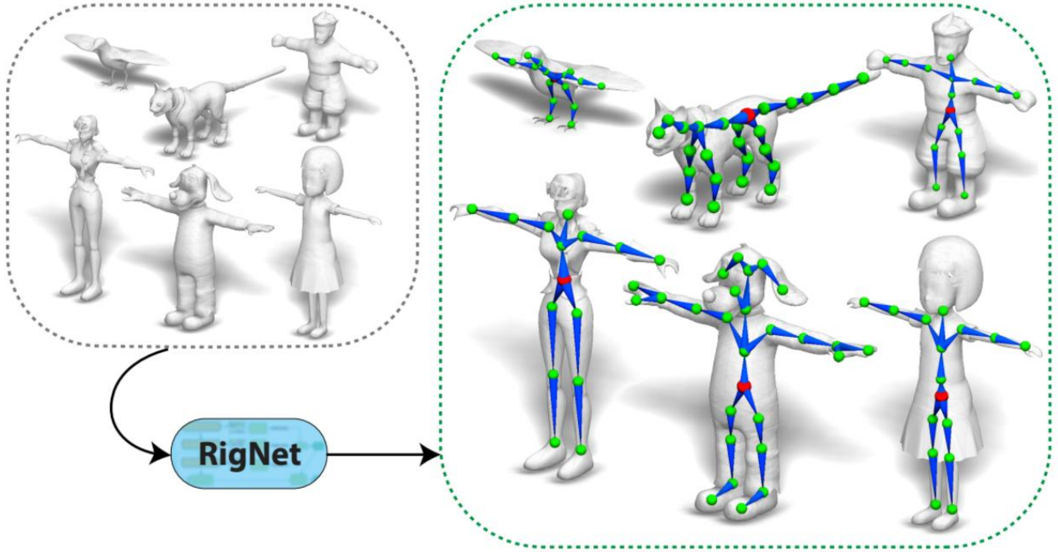
Linear blend skinning (LBS): $\mathbf{v}' = (\sum_{i=1}^{n} w_i T_i)\mathbf{v}$

Previous solutions

Manual rigging:

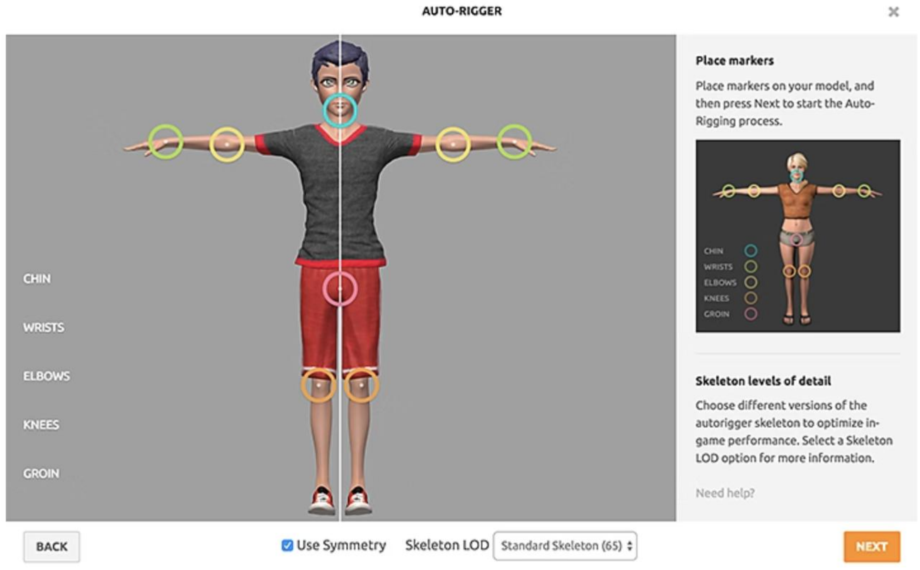
Manual rigging is time-consuming and requires significant expertise.

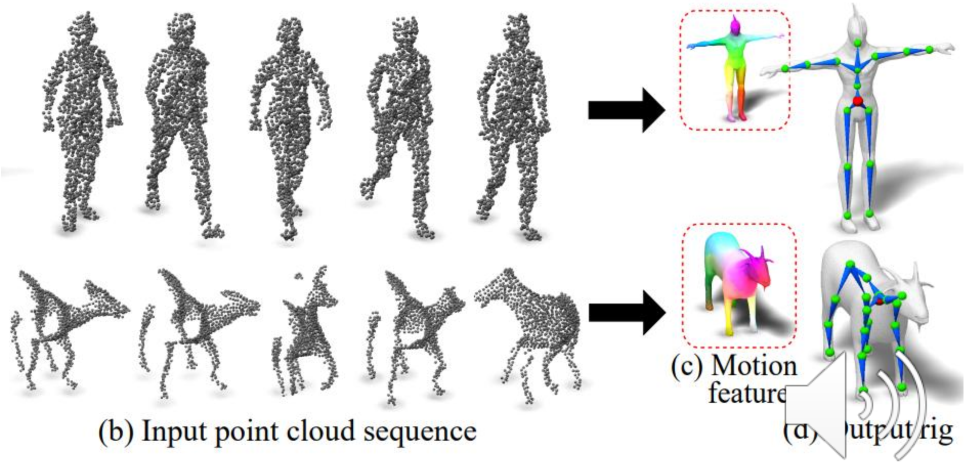




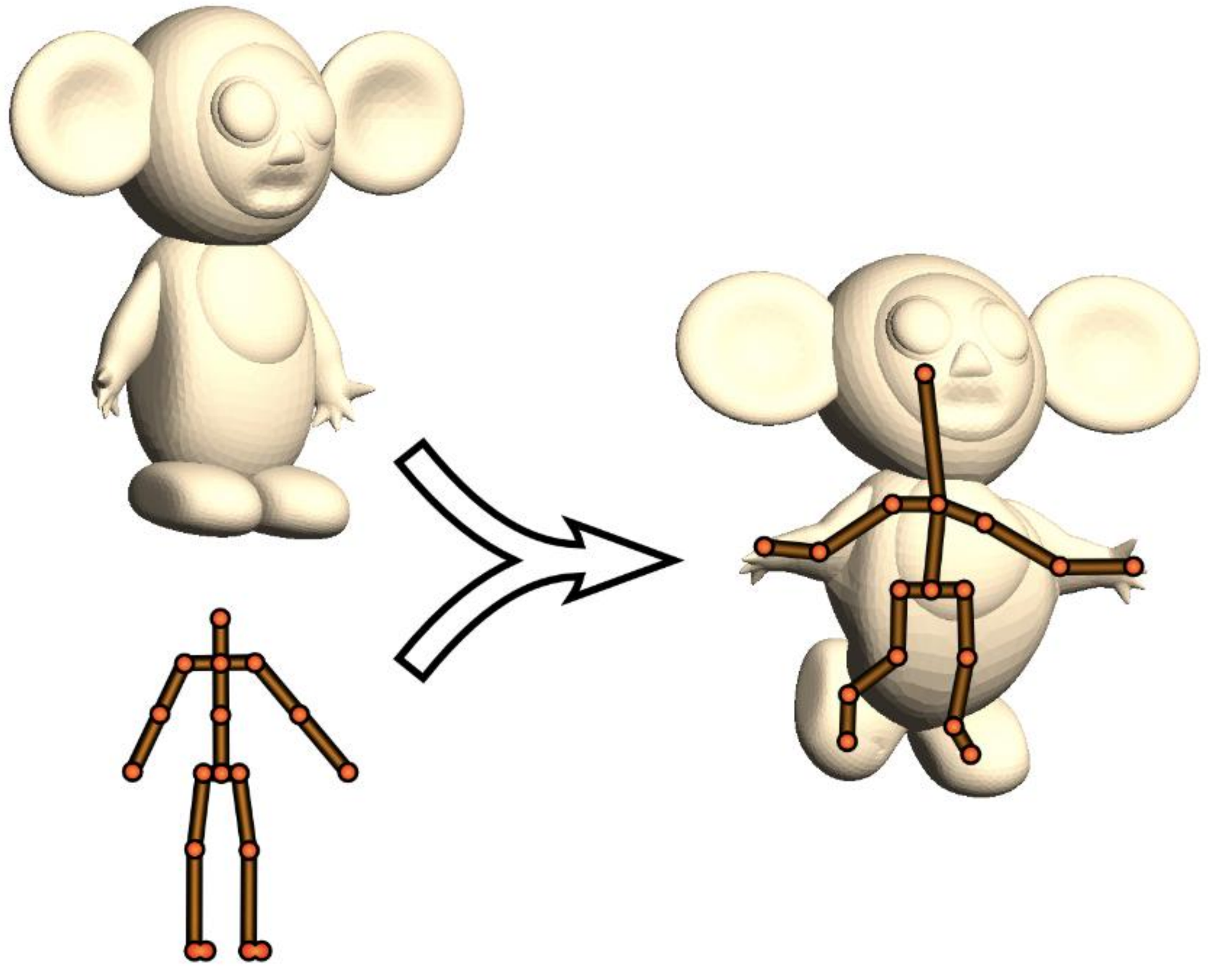
Automatic rigging:

1. Template-based
2. Template-free
3. Rely on additional inputs





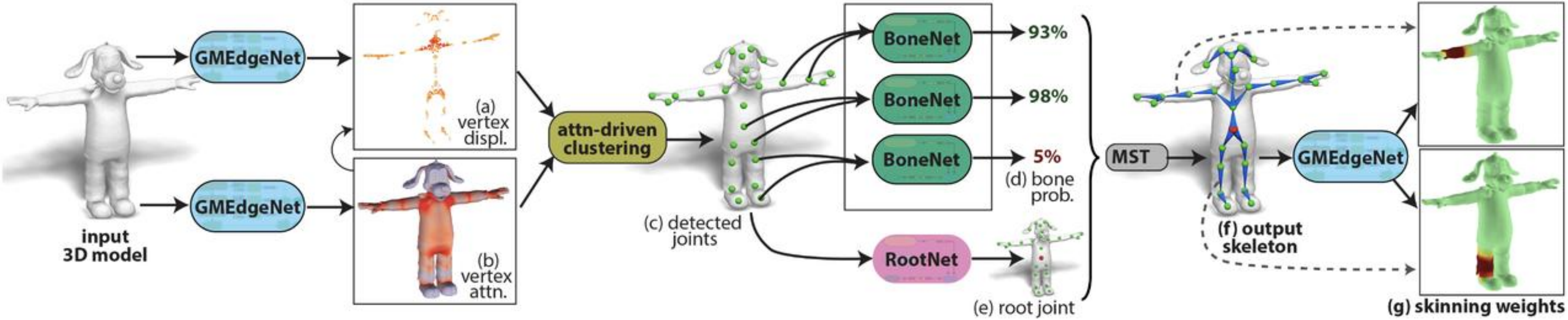
Previous solutions: template-based



- Rely on predefined templates.
- Fit a predefined skeleton template to the 3D model by minimizing the fitting cost.
- Difficult to generalize across diverse categories.

Automatic Rigging and Animation of 3D Characters (**Pinocchio**), Baran et al., Siggraph 2007

Previous solutions: template-free



- Strong assumption that input shapes maintain a consistent upright and front-facing orientation.

- Difficult to scale up.

- Introduce a small dataset with less than 3k models.

RigNet: Neural Rigging for Articulated Characters, Xu et al., Siggraph 2020

Previous solutions: Summary

- the lack of a **large-scale**, **diverse** dataset for training generalizable models.

- the need for an effective framework capable of handling **complex mesh topologies**, accommodating **varying skeleton structures.**

Our solution: MagicArticulate

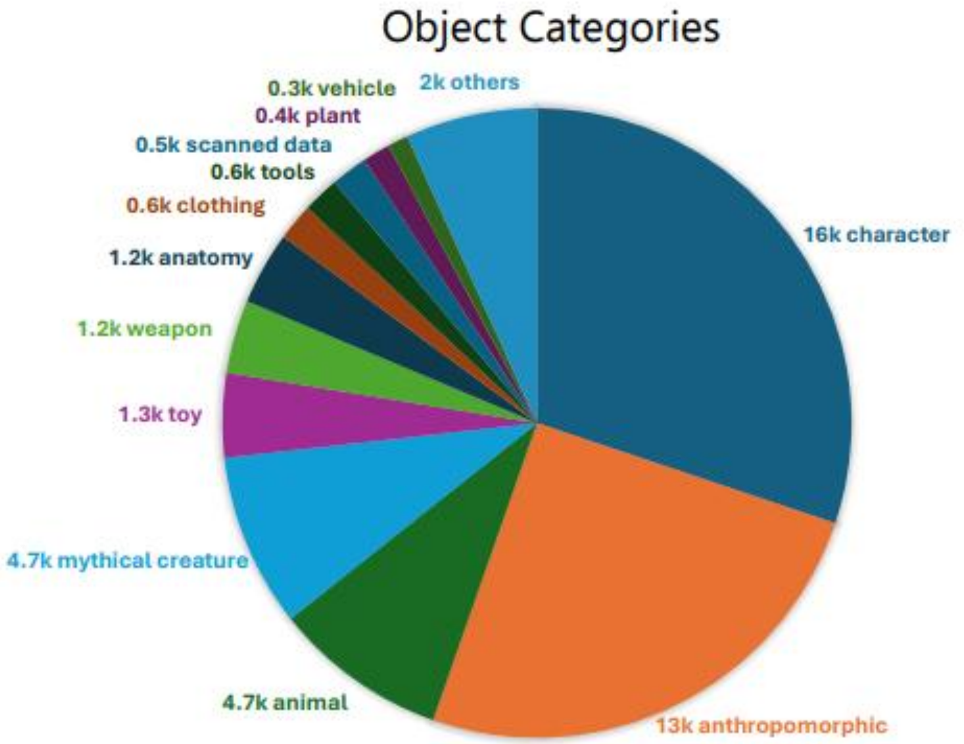
- Introduce **Articulation-XL**, a large-scale dataset containing over 33k 3D models with high-quality articulation annotations.

- Formulate skeleton generation as a **sequence modeling problem**.

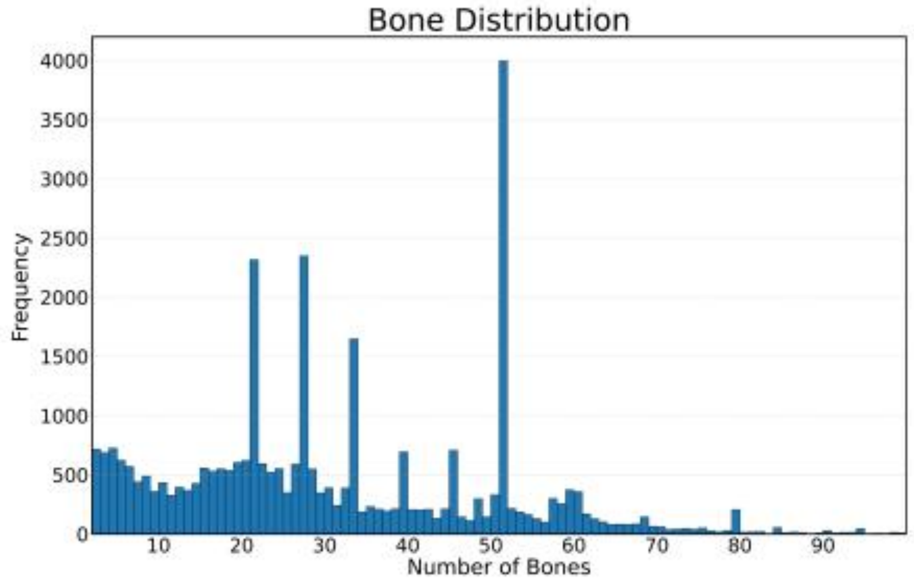
- Predict skinning weights using a **functional diffusion process**.

Dataset: Articulation-XL

(a) **Word cloud of Articulation-XL categories.**



(b) **Breakdown of Articulation-XL categories.**



(c) **Bone number distributions of Articulation-XL.**

Articulation-XL2.0 with over 48K data has been open sourced.
roughly 16× larger than the RigNet dataset.

Dataset: Articulation-XL

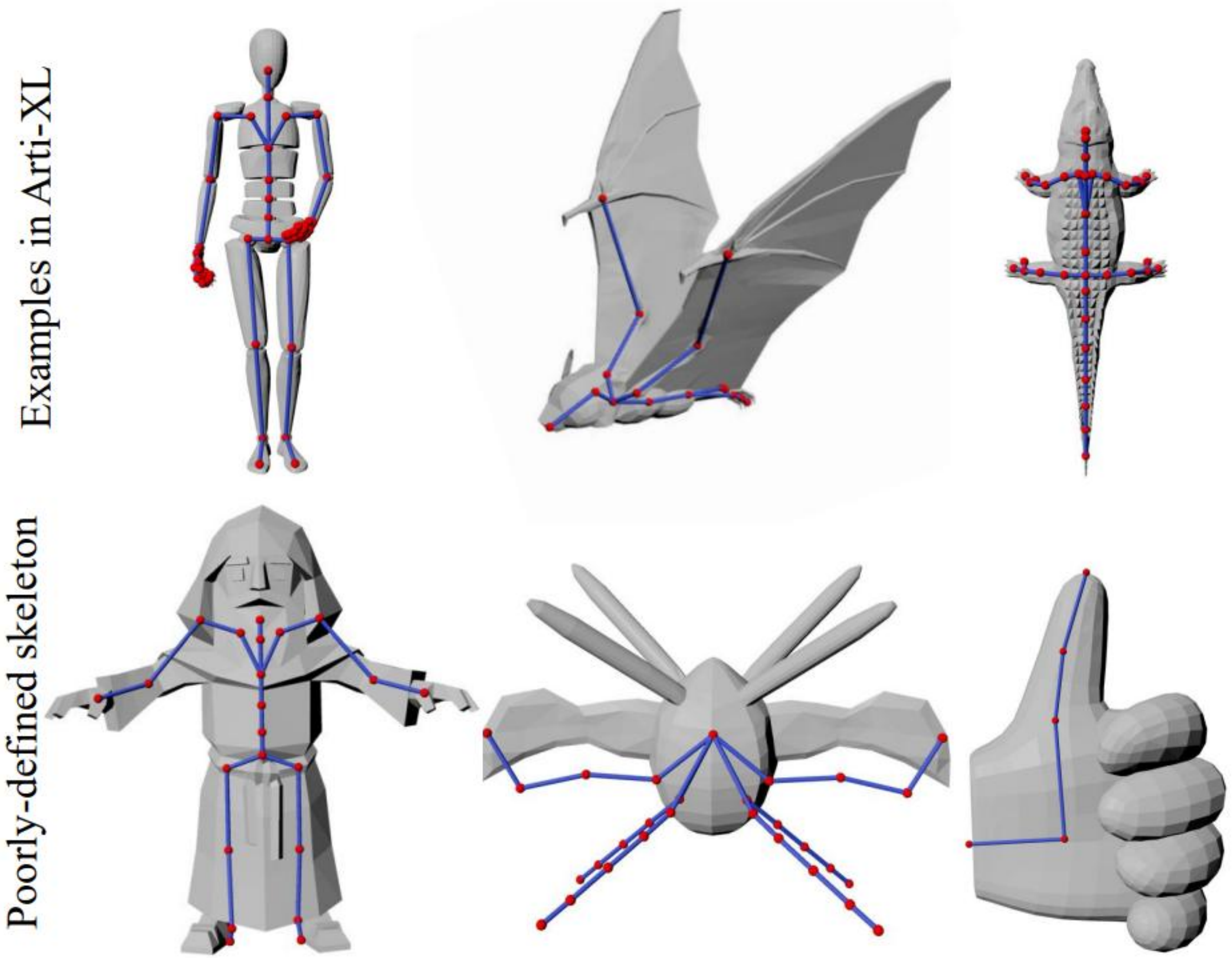
1. Initial data collection (glb, fbx, dae, etc).
2. VLM-based filtering and manual review.
3. Category label annotation.

Table 1. **Data statistics.**

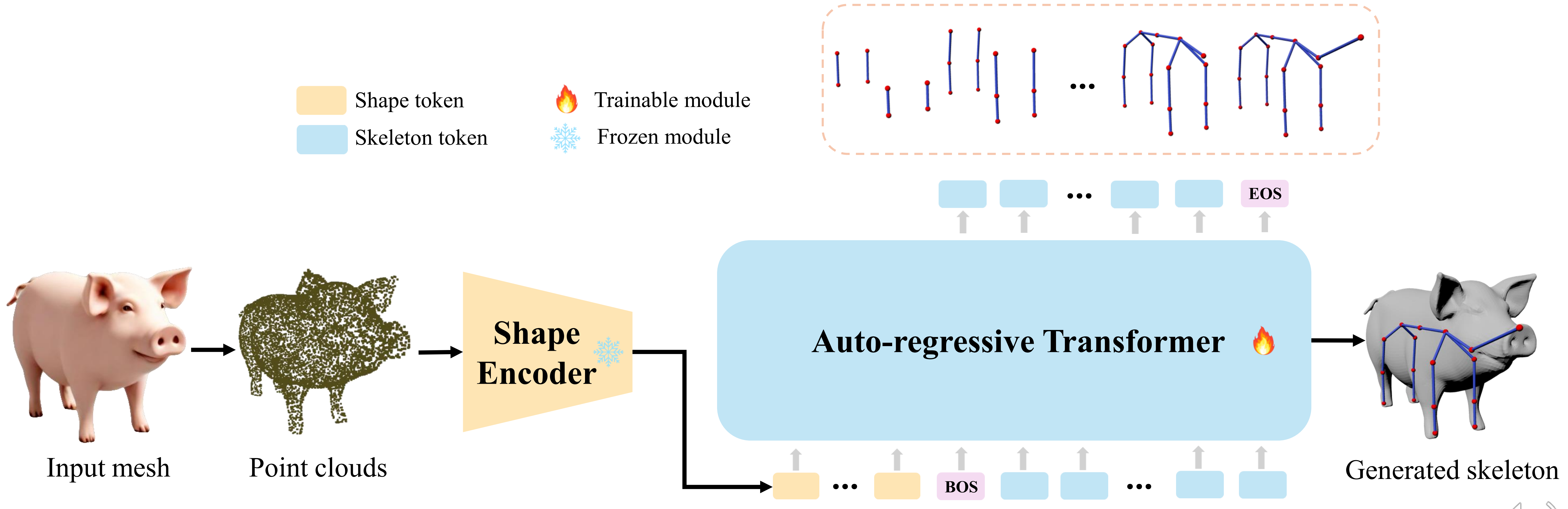
Source	All 3D data	with rigging	high quality rigging	low quality rigging
GitHub	2.08M	64K	42K	22K
Objaverse1.0	0.89M	10K	6K	4K
Sum	2.97M	74K	48K	26K

Articulation-XL2.0, the data with rigging has been deduplicated (over 150K).

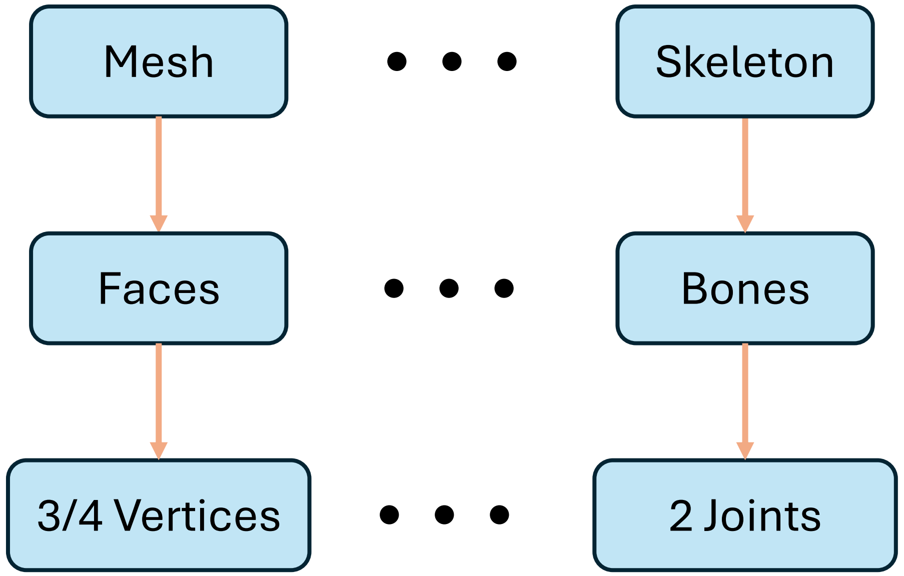
Dataset: some examples



Auto-regressive skeleton generation



Skeleton sequence modeling



Modeling skeleton as a sequence of bones.

Skeleton tokenization: sequence of bones

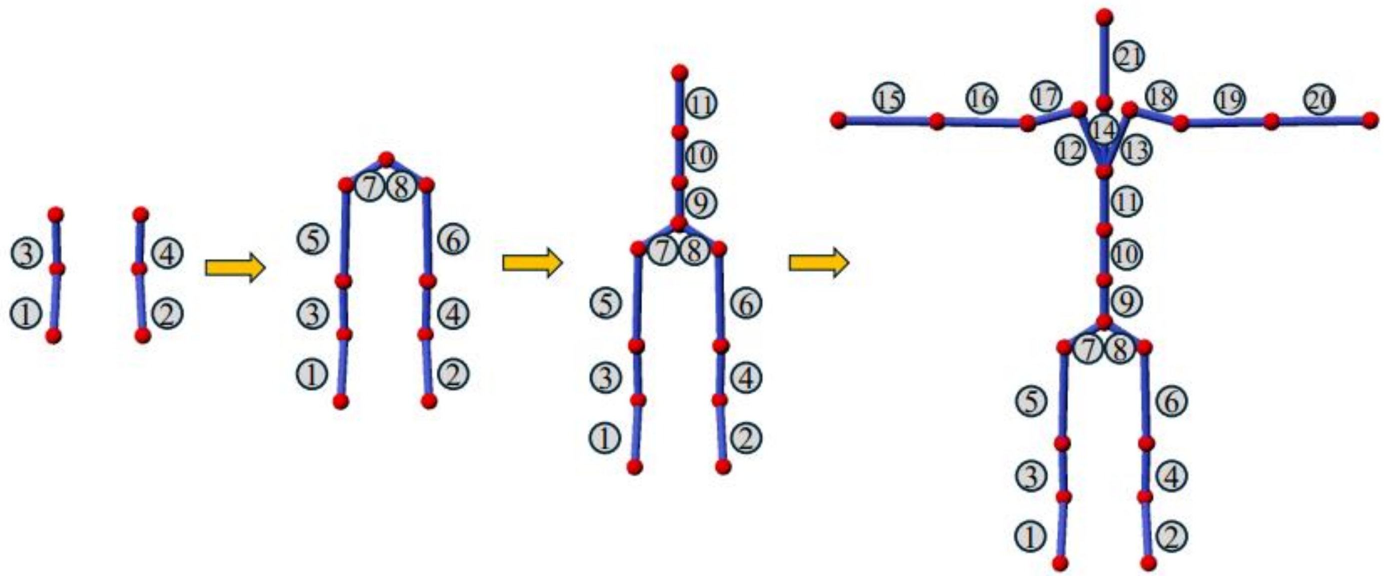
B1 = (x1, y1, z1, x2, y2, z2)

B2 = (x2, y2, z2, x3, y3, z3)

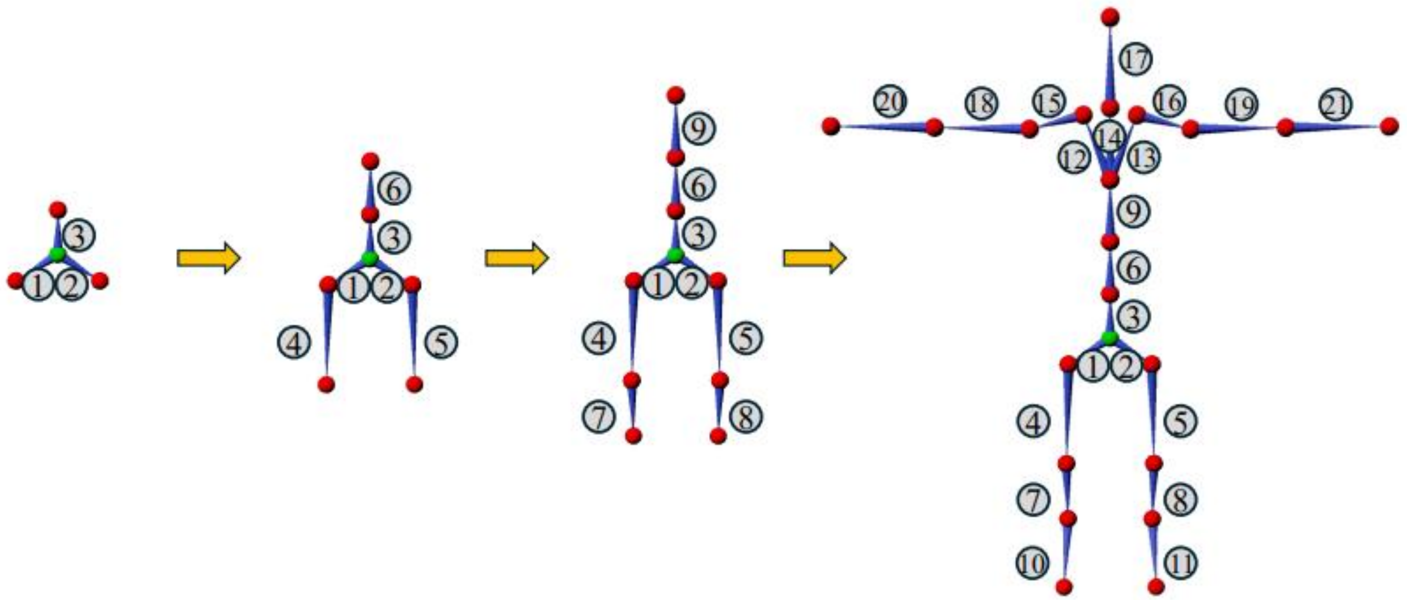
normalization --> discretization --> 6b sequence

How to sort this sequence?

Sequence ordering



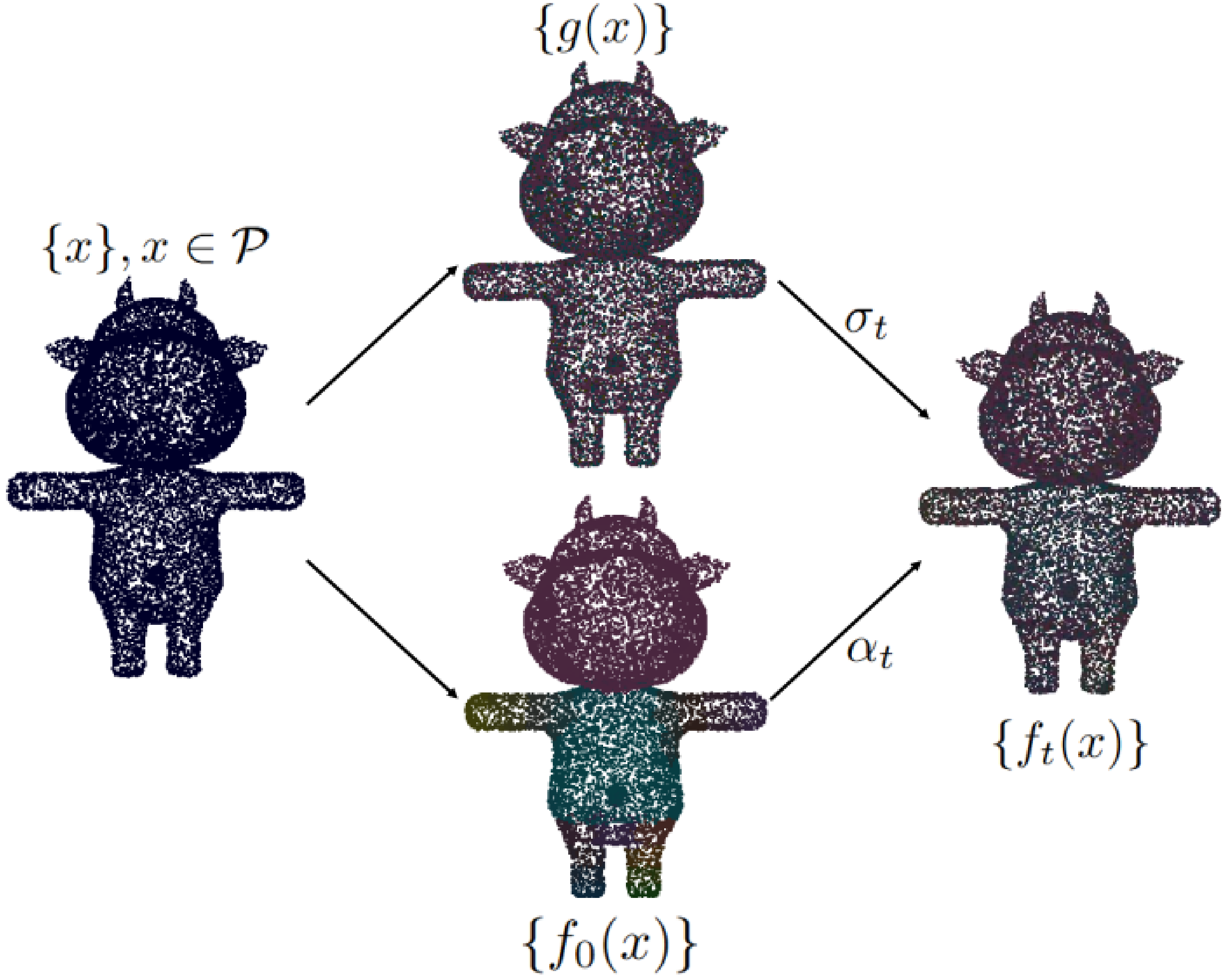
Spatial sequence ordering



Hierarchical sequence ordering

$$\mathcal{L}_{pred} = \mathrm{CE}(\mathbf{T}, \hat{\mathbf{T}})$$

Skinning weight prediction: functional diffusion

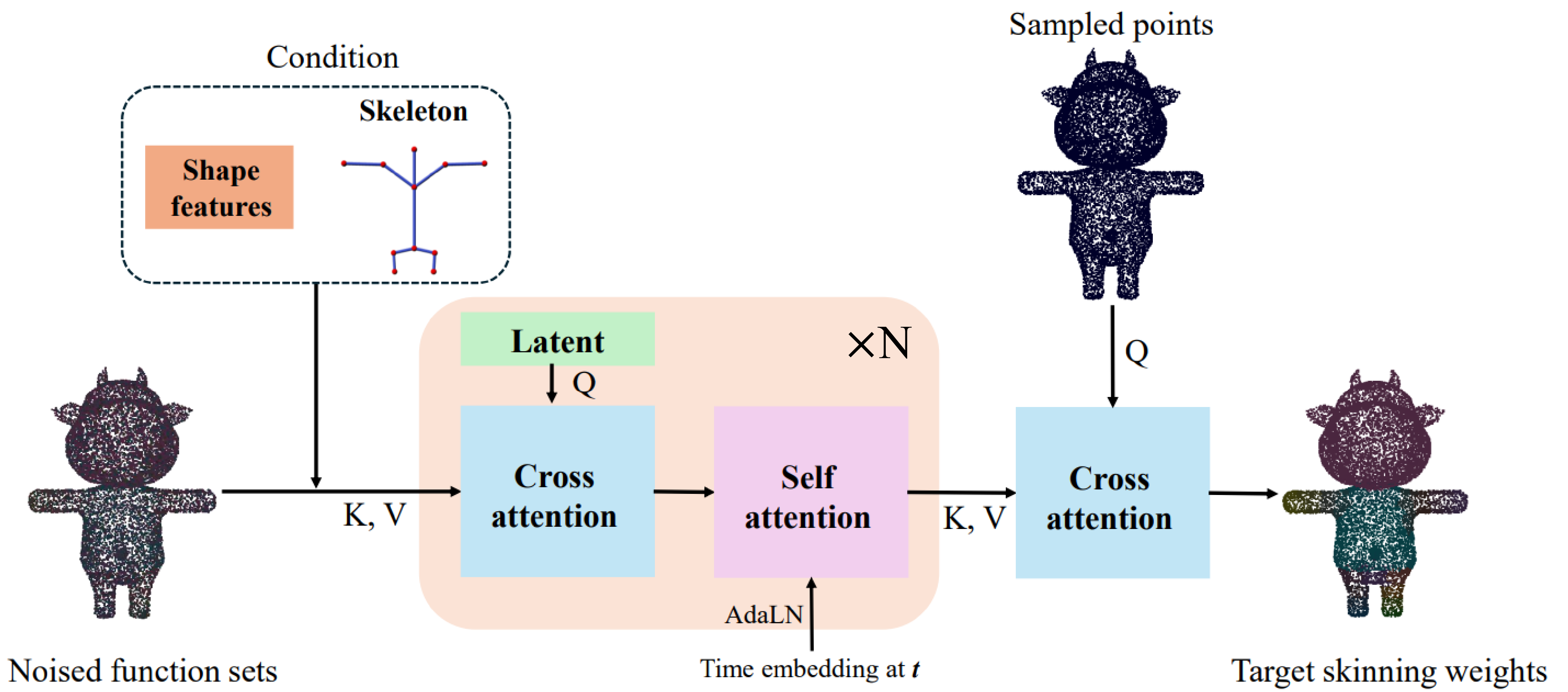


$$f_0 : \mathcal{X} \to \mathcal{Y}.$$

$$f_t(x) = \alpha_t \cdot f_0(x) + \sigma_t \cdot g(x), \quad t \in [0, 1]$$

$$D_\theta[f_t, t](x) \approx f_0(x).$$

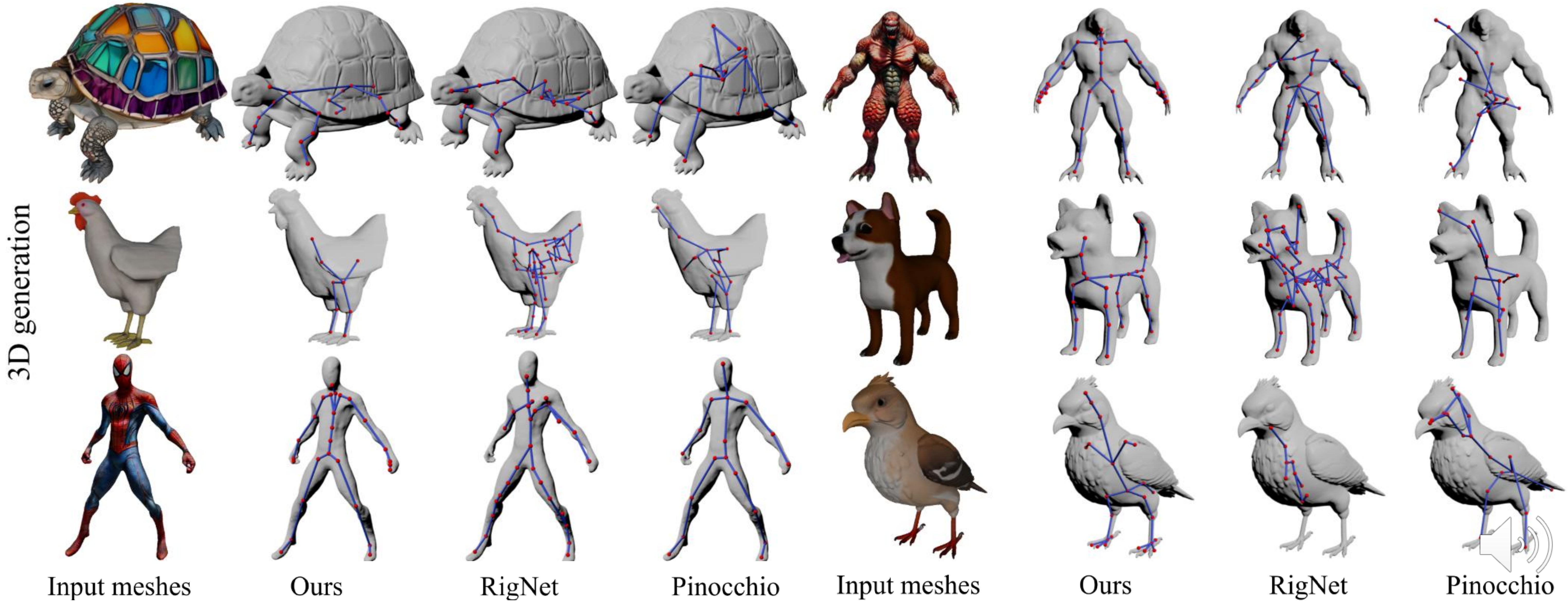
Skinning weight prediction



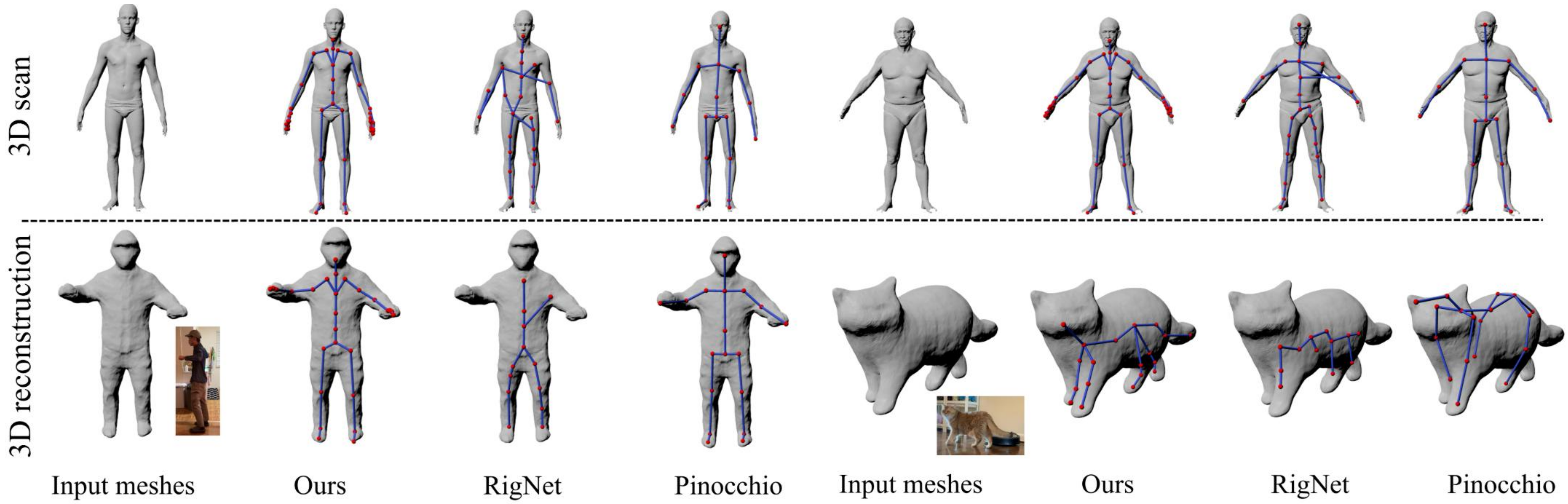
$$f : \mathcal{P} \rightarrow (\mathcal{W} - \mathcal{G})$$

$$\mathcal{L}_{denoise} = \left\| D_\theta \left(\{x, f_t(x)\}, t \right) - f_0(x) \right\|_2^2, \quad x \in \mathcal{P}.$$

Skeleton generation results



Skeleton generation results

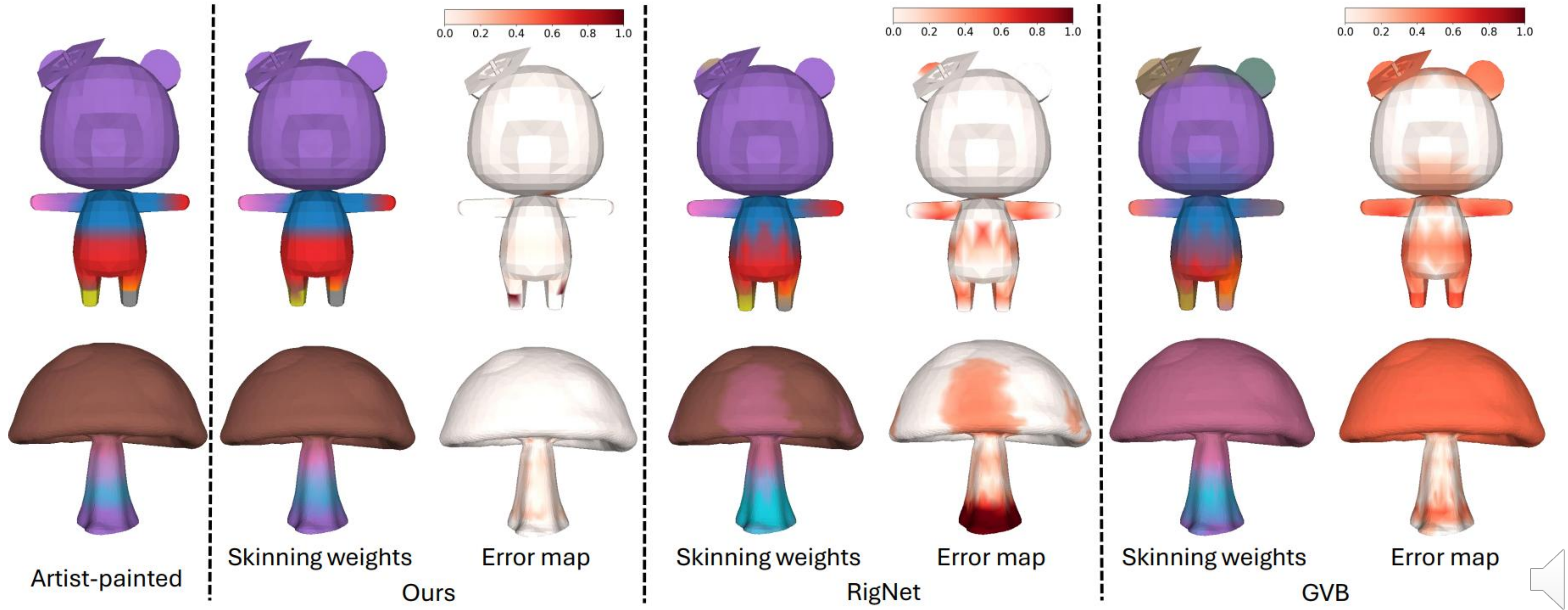


Skeleton generation results

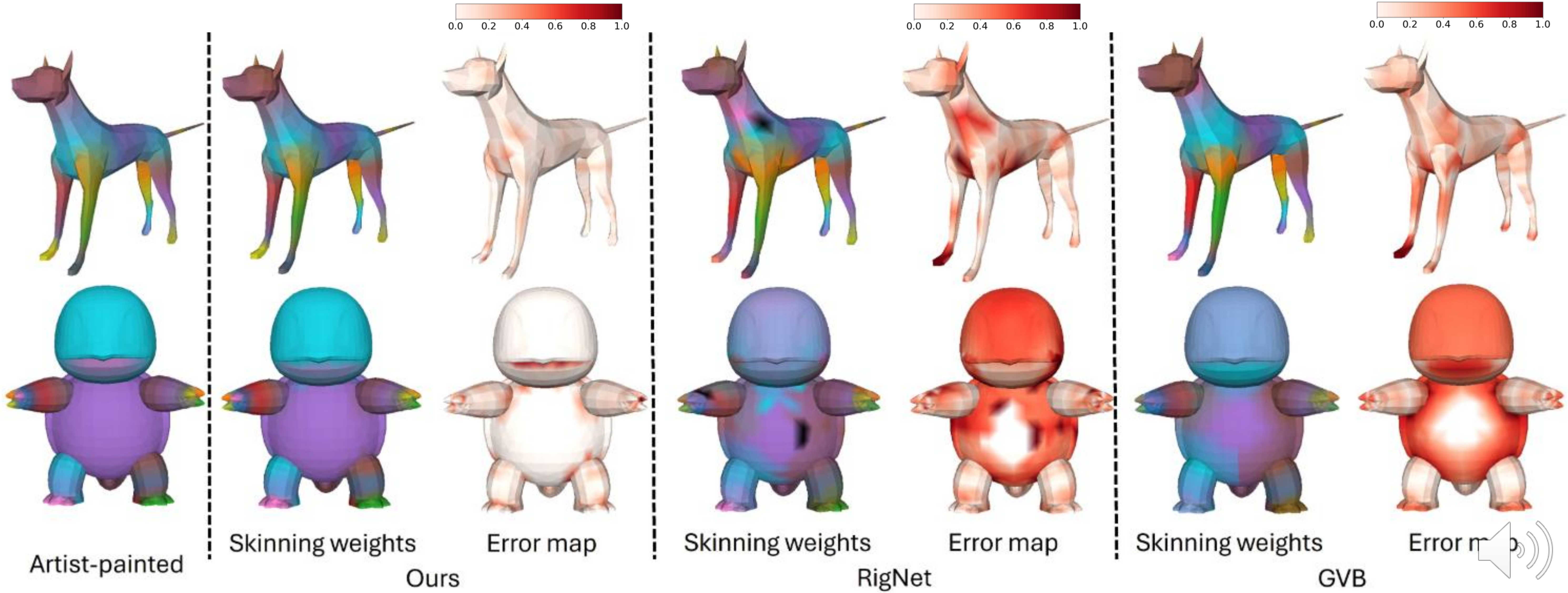
These Chamfer Distance-based metrics measure the spatial alignment between generated and ground truth skeletons. Lower is better.

	Dataset	CD-J2J	CD-J2B	CD-B2B
Pinocchio	*ModelsRes.*	6.852	4.824	4.089
RigNet		4.143	2.961	2.675
Ours-hier		<u>3.654</u>	<u>2.775</u>	<u>2.412</u>
Ours-spatial		**3.343**	**2.455**	**2.140**
Pinocchio	*Arti-XL*	8.360	6.677	5.689
RigNet		7.478	5.892	4.932
Ours-hier		<u>3.025</u>	<u>2.408</u>	<u>2.083</u>
Ours-spatial		**2.586**	**1.959**	**1.661**

Skinning weight prediction results

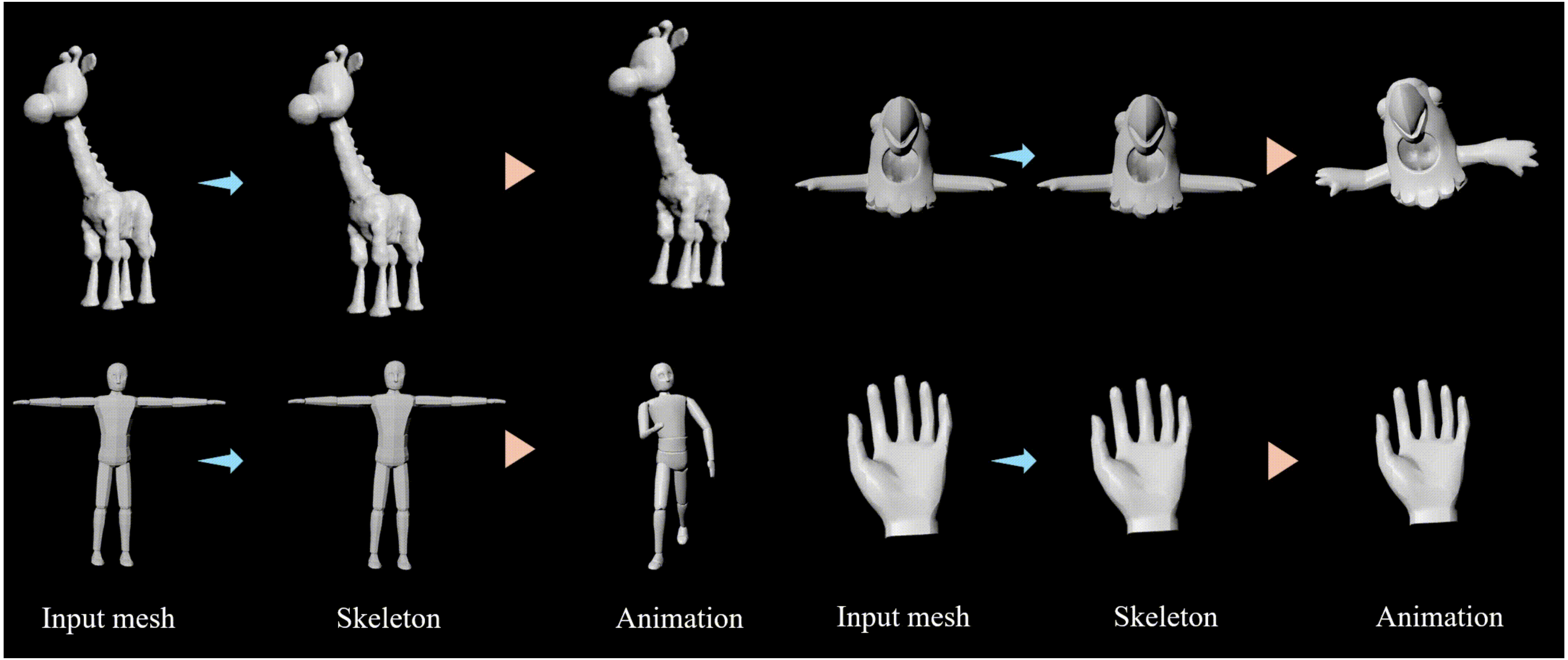


Skinning weight prediction results



Skinning weight prediction results

	Dataset	Precision	Recall	avg L1	avg Deformation
GVB		69.3%	79.2%	0.687	0.0067
RigNet	*ModelsResource*	77.1%	**83.5%**	0.464	0.0054
Ours		**82.1%**	81.6%	**0.398**	**0.0039**
GVB		75.7%	68.3%	0.724	0.0095
RigNet	*Articulation-XL*	72.4%	71.1%	0.698	0.0091
Ours		**80.7%**	**77.2%**	**0.337**	**0.0050**



However, animations still require manual efforts...

Rigging issues in MagicArticulate

1. Limited generalization to diverse pose inputs.

2. Skeleton sequence modeling can be more efficient.

3. Functional diffusion exhibits poor cross-dataset generalization and suffers from slow inference.

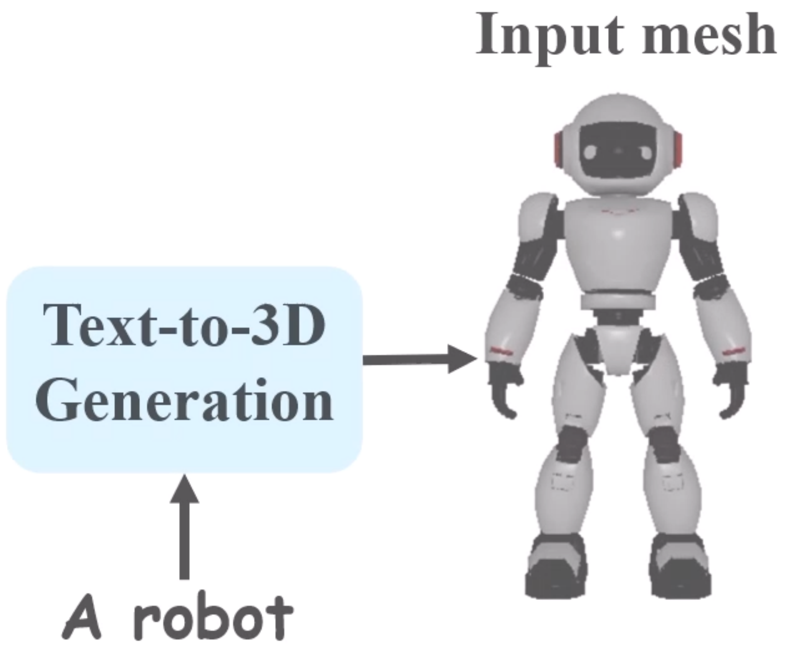
Automatic rigging and animation

Puppeteer: Rig and Animate Your 3D Models

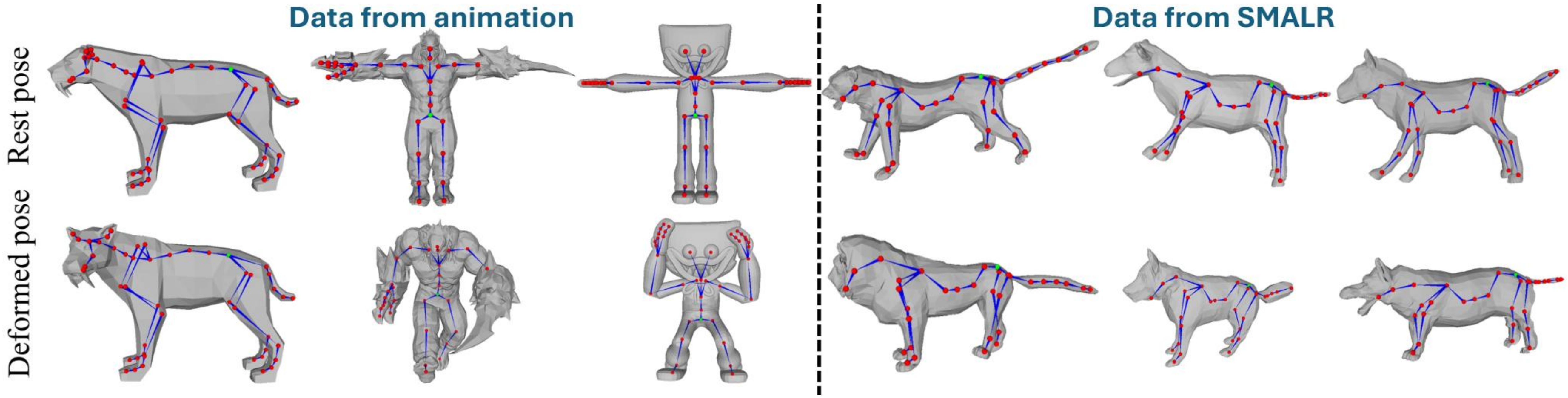
arXiv 2025

Chaoyue Song[1,2], Xiu Li[2], Fan Yang[1], Zhongcong Xu[2], Jiacheng Wei[1],
Fayao Liu[3], Jiashi Feng[2], Guosheng Lin[1*], Jianfeng Zhang[2*]
(* Corresponding authors)
[1]Nanyang Technological University, [2]Bytedance Seed, [3]A*STAR

Pipeline

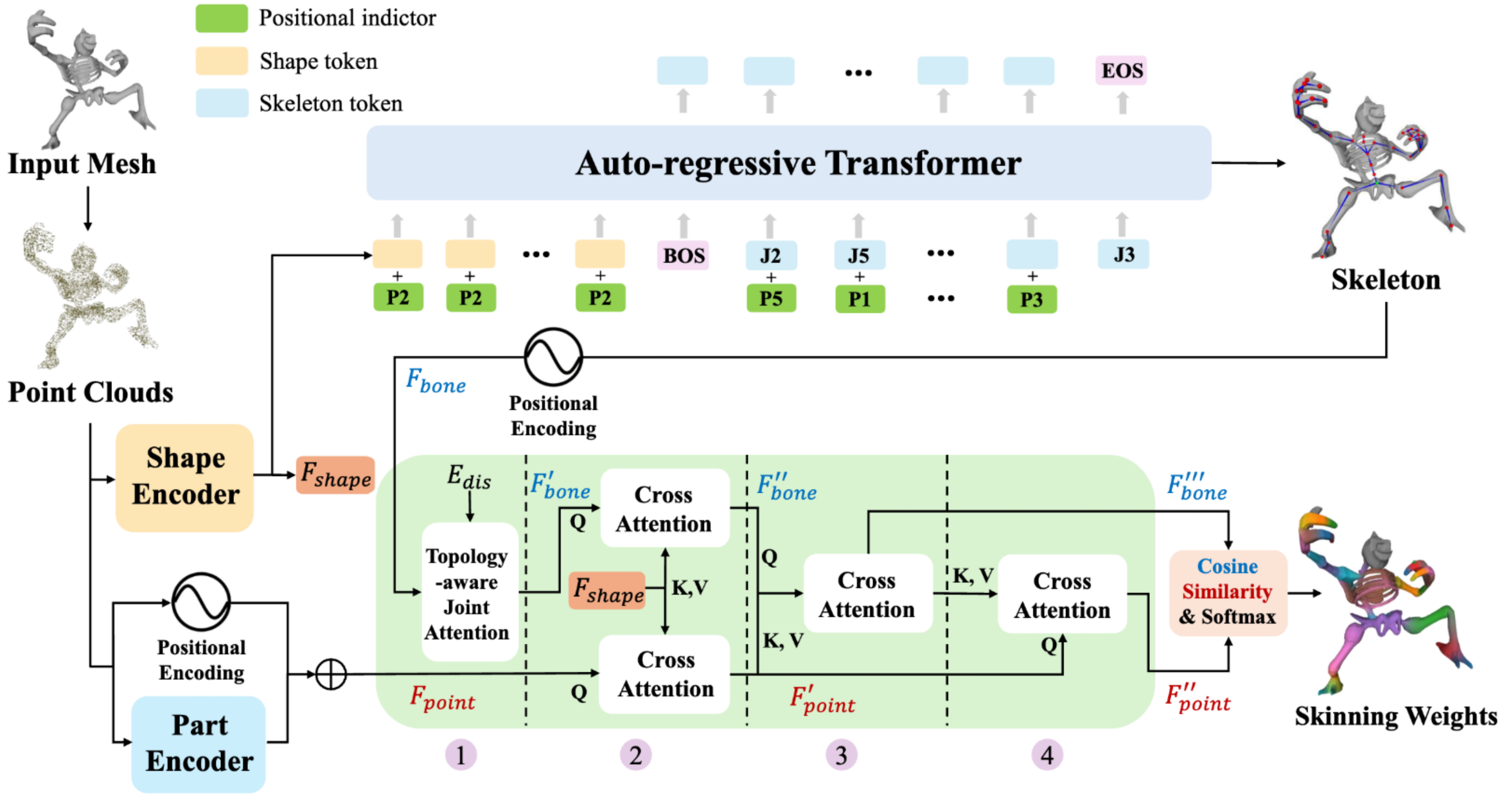


Dataset expansion

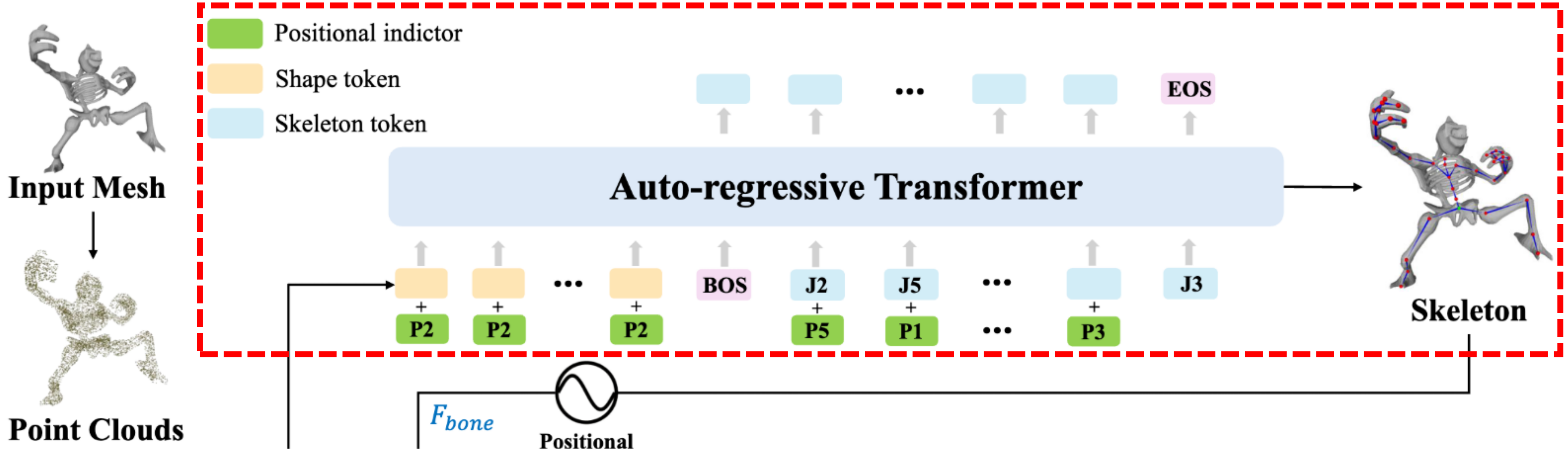


main set (48K) + diverse-pose subset (11.4K) = 59.4K

Automatic rigging



Automatic rigging: skeleton



Bone-based (6b):

$$[(x_0, y_0, z_0, x_1, y_1, z_1), (x_1, y_1, z_1, x_2, y_2, z_2), \dots, (x_{j-2}, y_{j-2}, z_{j-2}, x_{j-1}, y_{j-1}, z_{j-1})]$$

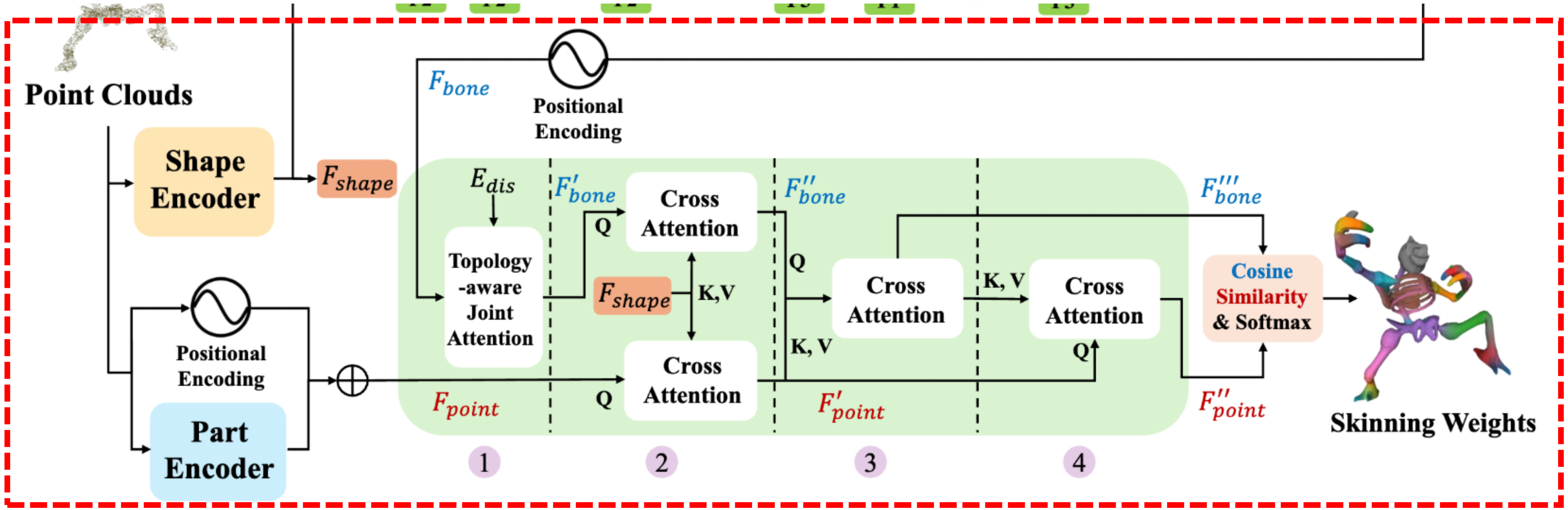
$$\mathbf{T} = [\mathbf{T}_{shape}, \mathbf{T}_{skel}] + \mathbf{P} = [\mathbf{T}_{shape} + \mathbf{p}_0, \mathbf{T}^0_{skel} + \mathbf{p}_1, \dots, \mathbf{T}^{j-2}_{skel} + \mathbf{p}_{j-1}, \mathbf{T}^{j-1}_{skel}]$$

Joint-based (4j): 4j < 6b whenever j > 3

$$[(x_0, y_0, z_0, p_0), (x_1, y_1, z_1, p_1), \dots, (x_{j-1}, y_{j-1}, z_{j-1}, p_{j-1})]$$

Automatic rigging: skinning weights

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{E}_{dis}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \lambda \mathbf{E}_{dis}\right)\mathbf{V}$$



Video-guided 3D animation

Input: rigged model, video $V = \{\mathbf{I}_0, \mathbf{I}_1, ..., \mathbf{I}_{n-1}\}$

For each frame $i \in \{1, 2, ..., n-1\}$

we optimize root motion $(\mathbf{Q}^i_{root}, \mathbf{T}^i_{root})$

joint-specific rotation $Q^i_{joint} = \{\mathbf{Q}^i_0, \mathbf{Q}^i_1, ..., \mathbf{Q}^i_{j-1}\}$

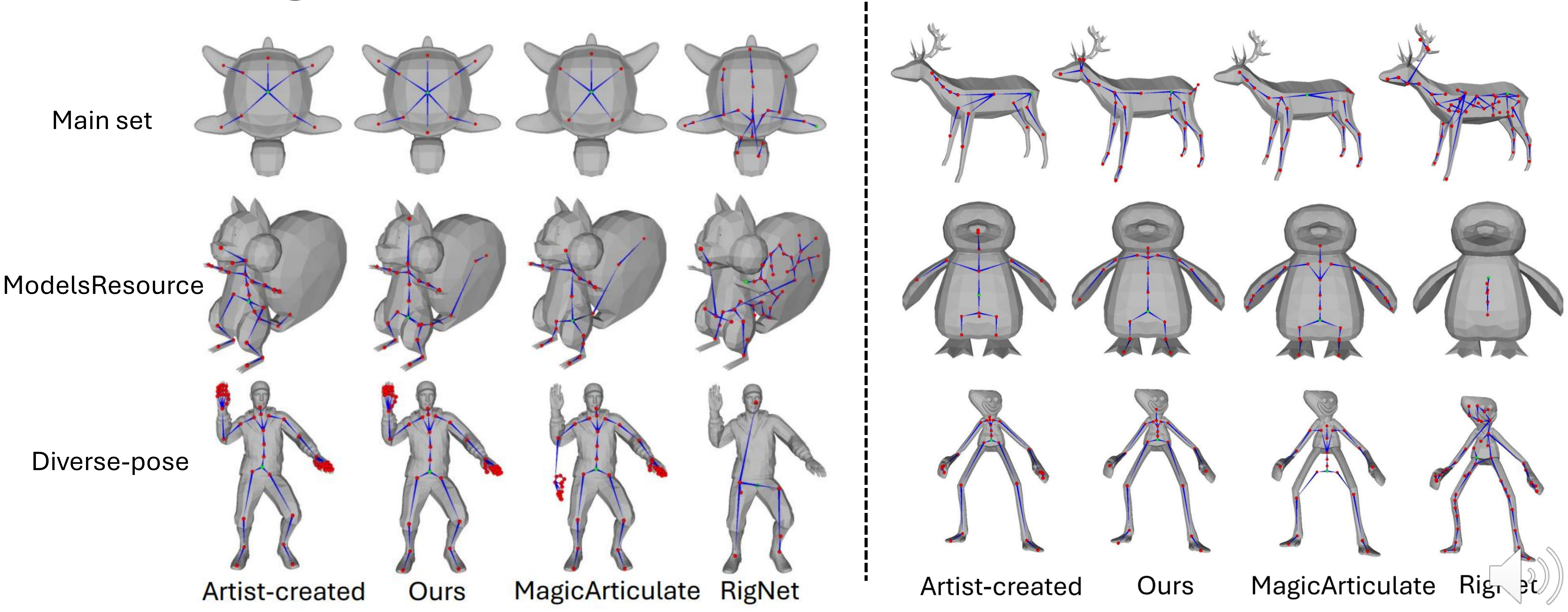
$$\mathcal{L} = \underbrace{(\mathcal{L}_{rgb} + \mathcal{L}_{mask} + \mathcal{L}_{flow} + \mathcal{L}_{depth})}_{\text{rendering losses}} + \underbrace{(\mathcal{L}_{joint\_track} + \mathcal{L}_{vertex\_track})}_{\text{tracking losses}} + \mathcal{L}_{reg}.$$

Experiments

- All dataset: main set (48K) + diverse-pose subset (11.4K)
- For training: main set (46K) + diverse-pose subset (10.9K)
- For test:

1. 2K from main set
2. 500 from the diverse-pose subset (rest pose also unseen)
3. 270 from ModelsResource, upright, front-facing, for cross-dataset generalization

Skeleton generation results

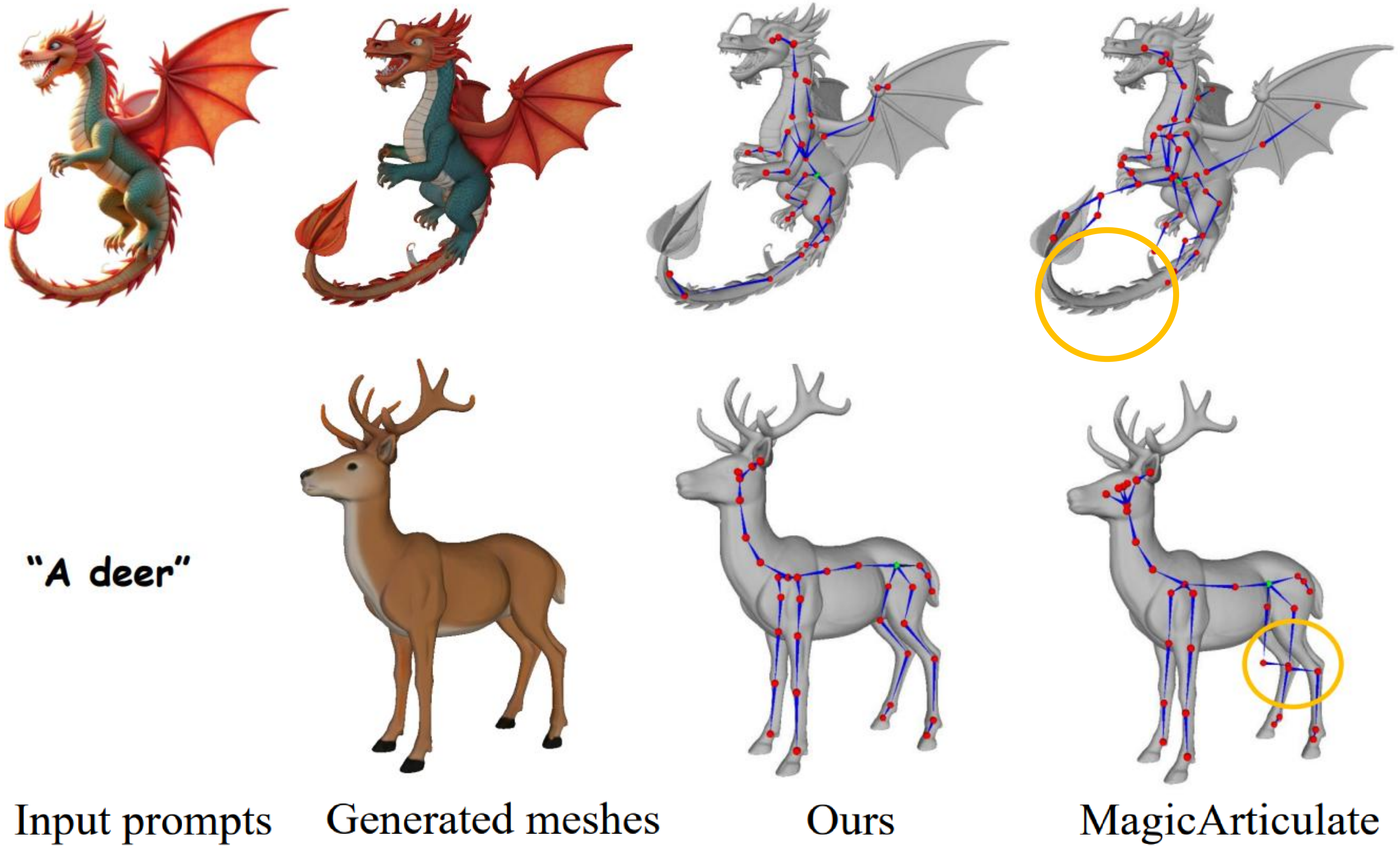


Skeleton generation results

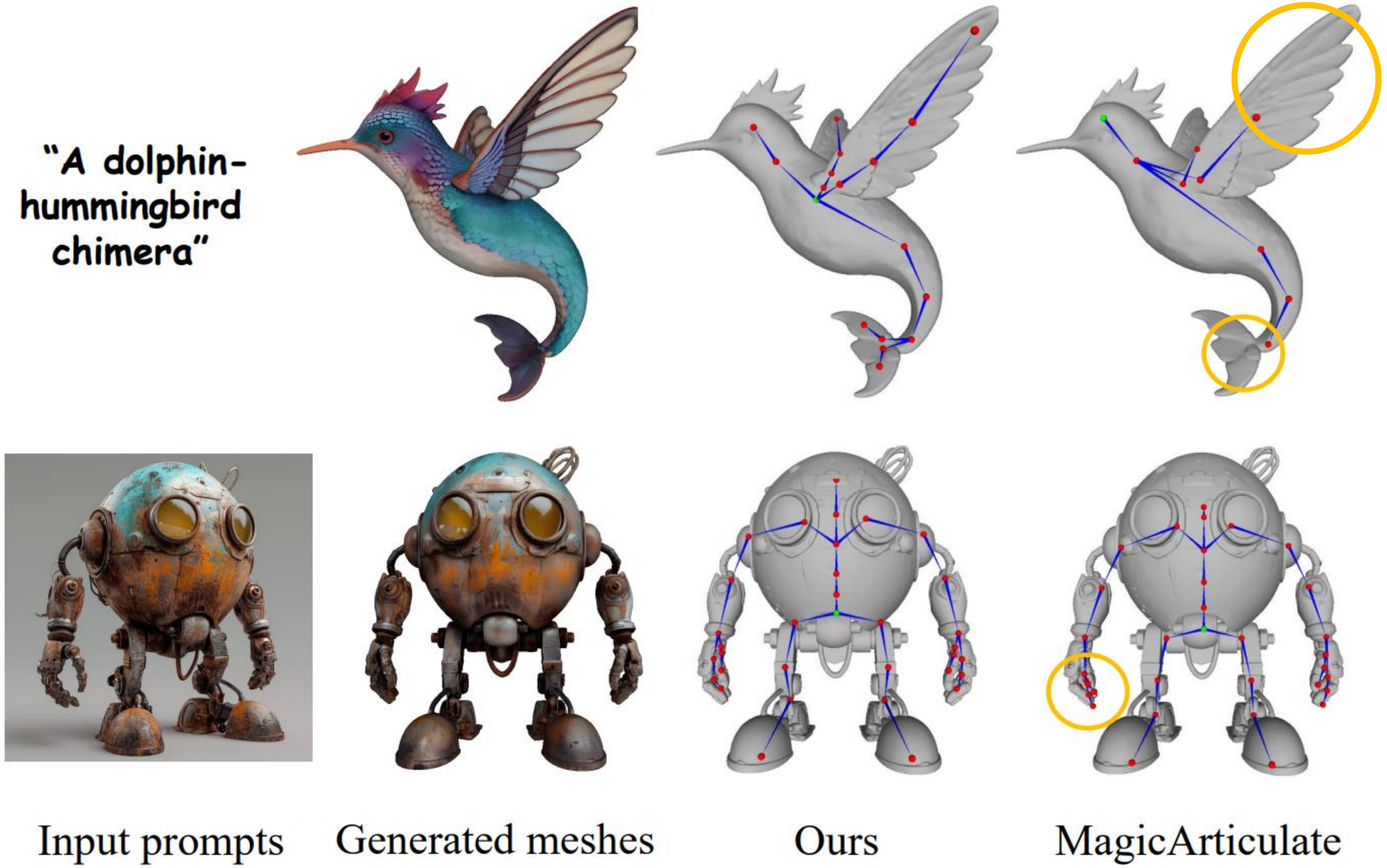
Method	Articulation-XL2.0			ModelsResource			Diverse-pose		
	J2J ↓	J2B ↓	B2B ↓	J2J ↓	J2B ↓	B2B ↓	J2J ↓	J2B ↓	B2B ↓
Pinocchio	8.324	6.612	5.485	6.852	4.824	4.089	7.967	6.411	5.149
RigNet	7.618	6.076	5.279	7.223	5.987	4.329	7.751	6.392	5.713
MagicArti.	3.264	2.503	2.123	4.114	3.137	2.693	4.376	3.456	2.955
UniRig	3.305	2.611	2.180	3.964	3.021	2.570	3.252	2.569	2.077
Ours	**3.033**	**2.300**	**1.923**	<u>3.841</u>	<u>2.881</u>	<u>2.475</u>	<u>3.212</u>	<u>2.542</u>	<u>2.027</u>
Ours*	<u>3.109</u>	<u>2.370</u>	<u>1.983</u>	**3.766**	**2.804**	**2.405**	**2.514**	**1.986**	**1.598**

Method	Pinocchio	RigNet	UniRig	MagicArticulate	Ours
Inference time	3.9s	4.5s	2.9s	2.4s	1.5s

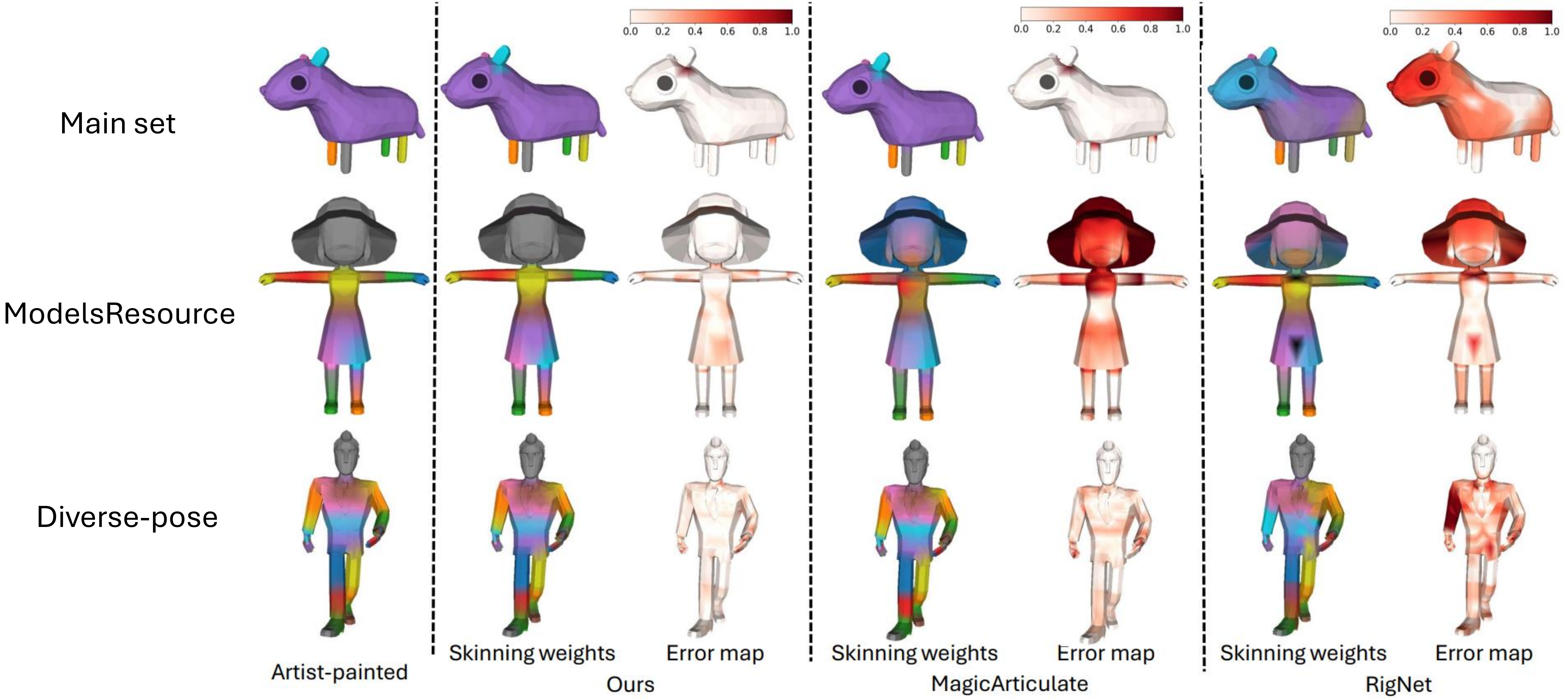
Skeleton results on AI-generated meshes



Skeleton results on AI-generated meshes



Skinning weight results

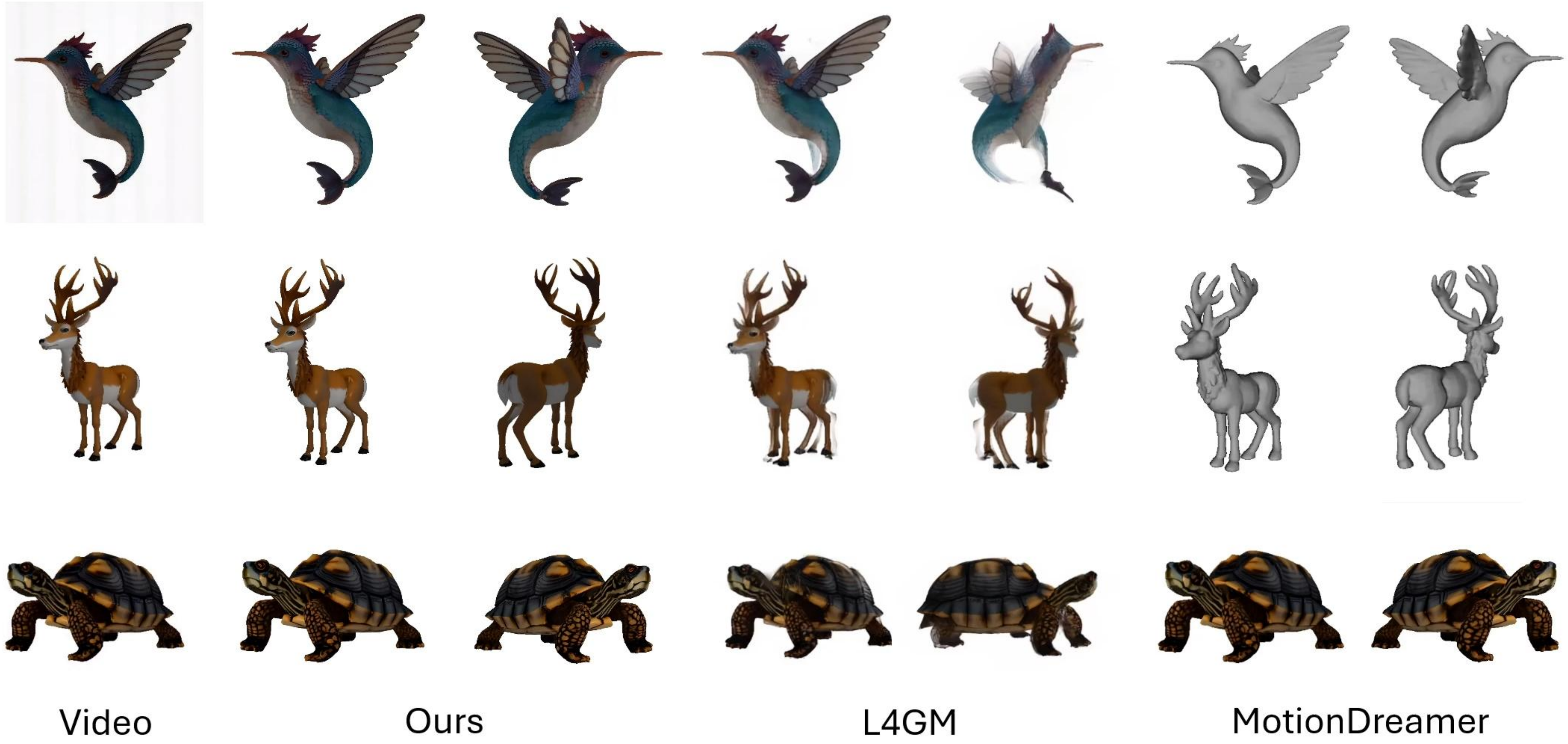


Skinning weight results

Method	Articulation-XL2.0			ModelsResource			Diverse-pose		
	Prec. ↑	Rec. ↑	L1 ↓	Prec. ↑	Rec. ↑	L1 ↓	Prec. ↑	Rec. ↑	L1 ↓
GVB	72.9%	65.5%	0.745	69.3%	79.2%	0.687	75.2%	64.9%	0.786
RigNet	73.7%	66.1%	0.729	65.7%	80.2%	0.707	74.7%	65.4%	0.746
MagicArti.	74.6%	71.3%	0.451	68.1%	80.7%	0.642	74.9%	68.4%	0.479
Ours	87.6%	**74.0%**	0.335	79.7%	**81.6%**	0.443	83.6%	72.2%	0.405
Ours*	**87.9%**	73.8%	**0.333**	**79.8%**	81.5%	**0.442**	**86.4%**	**72.8%**	**0.353**

Method	GVB	RigNet	MagicArticulate	Ours
Inference time	1.895s	0.056s	1.430s	0.032s

Animation results



Feed forward 3D animation

1. The animation optimization takes more than 20 minutes per object.

2. Rendering and tracking losses can cause ambiguity.

3. Require multi-view supervision.

Thanks!