FAMOS: Feed-Forward 3D Articulation Modeling from Sparse Observations

Kevin Qu[1,2], Tao Sun[1], Massimiliano Viola[1], Liyuan Zhu[1], Zhizhuo Zhou[1], Sayan Deb Sarkar[1], Konrad Schindler[2], and Iro Armeni[1]

[1] Stanford University
[2] ETH Zürich

Abstract. Modeling articulated objects from sparse monocular observations is challenging because each view provides only partial geometry and motion evidence. Existing feed-forward methods typically infer articulation from a single complete observation and therefore rely heavily on learned shape priors. We present FAMOS, a feed-forward model that jointly predicts movable-part segmentation and joint parameters from a sparse, unordered set of partial point clouds. FAMOS aggregates articulation cues across observations using a multi-state transformer with alternating state-wise and global attention, and introduces an observed articulation span objective that encourages the model to exploit motion evidence across the full input set. To improve training-data scale and diversity, we further develop a procedural generator that produces self-annotated articulated objects on the fly. Experiments show that FAMOS outperforms feed-forward and optimization-based baselines, with particularly large gains on zero-shot benchmarks.

1 Introduction

Articulated objects are ubiquitous in our lives, and modeling their movable parts is valuable for robotics [4, 8, 30], AR/VR [2, 12], and embodied AI [5, 22]. Existing approaches often require restrictive inputs: optimization-based methods rely on dense multi-view observations and costly per-object fitting [6, 18, 28], while video-based methods depend on temporally ordered sequences and tracking [3, 23, 29]. Feed-forward methods offer a scalable way by directly predicting articulation from images or 3D geometry [13–15, 17]. However, most operate on a single static observation. Without observing the object move, the model must infer its kinematic properties largely from learned shape priors, limiting generalization to unseen geometries and partial inputs. Moreover, most existing feed-forward approaches cannot exploit motion cues even when multiple articulation states are available. We propose FAMOS, a feed-forward method that jointly reasons over a sparse, unordered set of partial point clouds. A multi-state transformer alternates state-wise and global attention to aggregate geometric and articulation cues across observations. We further introduce an *observed articulation span* objective that supervises the motion range exhibited across the input states,

encouraging the model to leverage the complete observation set. The architecture supports a variable number of inputs, including a single observation. To increase training-data scale and diversity, we further introduce an on-the-fly procedural generator of self-annotated articulated assets. In summary, our contributions are: (i) a feed-forward method for articulated object modeling that jointly reasons over a sparse, unstructured set of observations. (ii) a procedural asset generator that synthesizes self-annotated articulated objects on the fly; and (iii) FAMOS consistently outperforms feed-forward and optimization-based baselines, with particularly large gains on zero-shot benchmarks, and also generalizes from synthetic training data to real-world captures.

2 Related Work

Articulated-object modeling has been approached through dense optimization [18, 21, 28], temporal interaction sequences [20, 24, 29], and feed-forward prediction [13, 14, 17]. Existing multi-observation methods typically assume complete geometry, temporal ordering, a fixed number of states, or prior knowledge of the part count [1, 7, 11, 16]. In contrast, FAMOS operates on sparse, unordered, variable-sized observation sets without requiring complete geometry or a known number of movable parts.

3 Method

Problem formulation. We consider an unordered set of S partial point clouds $\mathbf{X}_s = \{\mathbf{x}_{s,i}\}_{i=1}^{N}$ with each $N = 2048$ points in a shared coordinate frame. These can be obtained, e.g., from a 3D reconstruction model [26,27]. The number of observations may vary and the number of movable parts is unknown. The model predicts movable part segmentation together with per-part joint parameters: the joint type, an axis and origin for revolute joints, or a translation direction for prismatic joints. An overview of FAMOS is shown in Fig. 1.

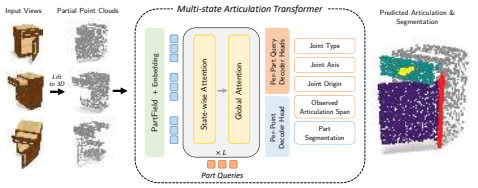

Fig. 1: FAMOS overview. Partial point clouds are processed by alternating state-wise and global attention and decoded into movable part segmentation and joint parameters.

Multi-State Articulation Transformer. Each observation is encoded independently using a frozen PartField [19] backbone and geometric embeddings. We use no observation-index embeddings, allowing the model to process a variable number of inputs. A set of learnable part queries represents candidate object parts. Point tokens and queries are refined by layers alternating between *state-wise attention*, which aggregates geometric context within each partial point cloud, and *global attention*, which jointly processes tokens from all observations and enables direct point-level interaction across them. Since a single query represents the same physical part in every state, the resulting

segmentation is cross-state consistent by construction. Lightweight heads decode the segmentation and joint parameters.

Observed articulation span. Joint type, axis, and origin are object-intrinsic properties and can, in principle, be inferred from a single observation. Supervising these quantities alone therefore does not require the model to integrate information across states. We introduce the *observed articulation span* $\Delta_p = \max_{s \in \mathcal{V}p} \theta s, p - \min_{s \in \mathcal{V}p} \theta s, p$, where $\mathcal{V}p$ denotes the observations in which part p is visible and $\theta s, p$ its joint value; the span loss is applied only when $|\mathcal{V}_p| \geq 2$. Unlike canonical joint limits, Δ_p depends on the configurations captured by the particular input set and therefore encourages reasoning across observations. As a difference of joint values, it is invariant to the unknown canonical rest state.

Procedural articulation data. To increase training-data scale and diversity, we introduce a parametric generator that assembles articulated objects from randomized analytic primitives. Dimensions, part counts, and joint placements vary across instances, while the resulting part segmentation and joint parameters are known exactly by construction. During training, new assets are synthesized on the fly, yielding a diverse stream of self-annotated articulated objects.

Training. Part queries are matched jointly to ground-truth parts across all observations using Hungarian matching and supervised for segmentation, joint type and parameters, and observed articulation span.

4 Experiments

Setup. We train on PartNet-Mobility [31], GRScenes [25], and procedurally generated assets. We evaluate on the PartNet-Mobility test split for in-domain performance and on ACD [9] and ArtiCraft-10K [32] as zero-shot benchmarks. We report movable-part segmentation F_1 at IoU 0.5 and motion-estimation F_1 under the MAO criterion [9,10], requiring the correct joint type, axis error below 5°, and origin error below 0.05 of the object scale. Unless stated otherwise, feed-forward methods receive two input states. As our closest feed-forward baseline, we re-train Particulate [14] on our data and extend it for multiple input point clouds. We additionally compare against the optimization-based ReArt [20] and ArtGS [21].

Main results. FAMOS outperforms all baselines in segmentation and motion estimation across the three benchmarks (Tab. 1). The gains over the multi-state

Table 1: Main results. Movable-part segmentation (Seg) and motion-estimation (MAO) F_1. † denotes re-trained on our data. S denotes the number of input states.

Method	S	**PartNet-M.**		**ACD**		**ArtiCraft**	
		Seg	MAO	Seg	MAO	Seg	MAO
ReArt [20]	4	58.2	33.3	22.5	12.3	44.1	34.7
ArtGS [21]	2	55.5	39.7	24.7	13.2	47.4	33.0
Particulate† [14]	2	93.6	90.3	43.7	36.7	76.6	68.5
Ours	2	**98.3**	**98.4**	**71.9**	**64.5**	**92.5**	**90.0**
Particulate† [14]	1	90.4	86.5	43.4	35.6	75.1	62.0
Ours	1	**95.2**	**90.7**	**68.5**	**59.4**	**89.9**	**83.0**

Table 2: Ablations on ACD. Left: model ablations (trained only on curated data). Right: effect of adding our procedurally generated assets to the training data.

Model components			Training data		
Variant	Seg	MAO	Setting	Seg	MAO
w/o Span Loss	52.3	41.6	Curated only	53.2	47.3
w/o Global Attention	46.5	39.2	+ Proc. pretrain	65.3	58.2
Full Model	**53.2**	**47.3**	+ Data mix	**71.9**	**64.5**

Fig. 2: Quantitative and qualitative results. *Left:* On ACD, prediction accuracy improves as more observations are provided, with the largest gain from one to two states. *Middle:* Qualitative results on benchmarks. *Right:* FAMOS generalizes to real-world captures reconstructed with VGGT-Ω [27] despite being trained only on synthetic data.

Particulate baseline are particularly pronounced on the two zero-shot benchmarks, while both methods perform closer to the ceiling on in-domain PartNet-Mobility. These results suggest that FAMOS can effectively leverage observed motion for more accurate predictions. In the single-state setting, FAMOS also consistently outperforms Particulate across all benchmarks, showing that it remains effective when only one observation is available. We show qualitative results in Fig. 2.

Ablations. Tab. 2 evaluates our model components and training data. Replacing global attention with register-based cross-state communication degrades performance, highlighting the importance of direct point-level interaction across observations. Removing the observed articulation span loss likewise reduces accuracy, showing that the auxiliary objective encourages effective aggregation of multi-state cues. Adding our procedurally generated training data consistently improves both segmentation and motion estimation. Pre-training only on generated assets already provides substantial gains, while continuing to mix generated and curated assets further improves performance. As shown in Fig. 2, performance also improves with additional observations, with the largest gain from one to two states, when explicit motion evidence first becomes available and can be leveraged by the model.

5 Conclusion

We presented FAMOS, a feed-forward method for articulated object modeling from sparse, unordered observations. By combining a multi-state transformer with an observed articulation span objective, FAMOS explicitly exploits motion cues across observations while supporting a variable number of inputs. A procedural asset generator further improves training-data scale and diversity. Across in-domain and zero-shot benchmarks, FAMOS consistently outperforms feed-forward and optimization-based baselines, and also generalizes to real-world captures.

References

1. Artykov, A., Ravaud, T., Sautier, C., Lepetit, V.: sim2art: Accurate articulated object modeling from a single video using synthetic training data only. arXiv preprint arXiv:2512.07698 (2025)
2. Behravan, M., Haghani, M., Gračanin, D.: Transcending dimensions using generative ai: Real-time 3d model generation in augmented reality. In: International conference on human-computer interaction. pp. 13–32. Springer (2025)
3. Delitzas, A., Zhang, C., Gavryushin, A., Di Mario, T., Sun, B., Dabral, R., Guibas, L., Theobalt, C., Pollefeys, M., Engelmann, F., Barath, D.: Reconstructing Functional 3D Scenes from Egocentric Interaction Videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
4. Gadre, S.Y., Ehsani, K., Song, S.: Act the part: Learning interaction strategies for articulated object part discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15752–15761 (October 2021)
5. Gao, X., Gong, R., Shu, T., Xie, X., Wang, S., Zhu, S.C.: Vrkitchen: an interactive 3d virtual environment for task-oriented learning (2019), `https://arxiv.org/abs/1903.05757`
6. Guo, J., Xin, Y., Liu, G., Xu, K., Liu, L., Hu, R.: Articulatedgs: Self-supervised digital twin modeling of articulated objects using 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27144–27153 (2025)
7. Gupta, A., Gu, W., Patil, O., Lee, J.K., Gopalan, N.: Pokenet: Learning kinematic models of articulated objects from human observations (2026), `https://arxiv.org/abs/2602.02741`
8. Hausman, K., Niekum, S., Osentoski, S.., Sukhatme, G.S.: Active articulation model estimation through interactive perception. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). pp. 3305–3312. IEEE (2015)
9. Iliash, D., Jiang, H., Zhang, Y., Savva, M., Chang, A.X.: S2o: Static to openable enhancement for articulated 3d objects. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6785–6795 (2026)
10. Iliash, D., Liu, J., Fokin, E., Wu, Q., Mahdavi-Amiri, A., Savva, M., Chang, A.X.: Artiverse: A diverse and physically grounded dataset for articulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
11. Jiang, Z., Hsu, C.C., Zhu, Y.: Ditto: Building digital twins of articulated objects from interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5616–5626 (2022)
12. Krichenbauer, M., Yamamoto, G., Taketom, T., Sandor, C., Kato, H.: Augmented reality versus virtual reality for 3d object manipulation. IEEE transactions on visualization and computer graphics **24**(2), 1038–1048 (2017)
13. Li, H., Xie, H., Xu, J., Wen, B., Hong, F., Liu, Z.: MonoArt: progressive structural reasoning for monocular articulated 3d reconstruction. arXiv preprint arXiv 2603.19231 (2026)
14. Li, R., Yao, Y., Zheng, C., Rupprecht, C., Lasenby, J., Wu, S., Vedaldi, A.: Particulate: Feed-forward 3d object articulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27708–27718 (2026)
15. Li, Z., Bai, X., Zhang, J., Wu, Z., Xu, C., Li, Y., Hou, C., Zhang, S.: Urdf-anything: Constructing articulated objects with 3d multimodal language model. Advances in Neural Information Processing Systems **38**, 94974–95002 (2026)

16. Li, Z., Zhang, C., Li, Z., Howard-Jenkins, H., Lv, Z., Geng, C., Wu, J., Newcombe, R., Engel, J., Dong, Z.: Art: Articulated reconstruction transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7468–7479 (2026)
17. Liu, J., Iliash, D., Chang, A., Savva, M., Mahdavi Amiri, A.: Singapo: Single image controlled generation of articulated parts in objects. In: International Conference on Learning Representations. vol. 2025, pp. 97511–97532 (2025)
18. Liu, J., Mahdavi-Amiri, A., Savva, M.: Paris: Part-level reconstruction and motion analysis for articulated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 352–363 (2023)
19. Liu, M., Uy, M.A., Xiang, D., Su, H., Fidler, S., Sharp, N., Gao, J.: Partfield: Learning 3d feature fields for part segmentation and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9704–9715 (2025)
20. Liu, S., Gupta, S., Wang, S.: Building rearticulable models for arbitrary 3d objects from 4d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21138–21147 (2023)
21. Liu, Y., Jia, B., Lu, R., Ni, J., Zhu, S.C., Huang, S.: Artgs: Building interactable replicas of complex articulated objects via gaussian splatting. ICLR (2025)
22. Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
23. Peng, W., Lv, J., Lu, C., Savva, M.: itaco: Interactable digital twins of articulated objects from casually captured rgbd videos. 3DV (2026)
24. Song, C., Wei, J., Foo, C.S., Lin, G., Liu, F.: Reacto: Reconstructing articulated objects from a single video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5384–5395 (2024)
25. Wang, H., Chen, J., Huang, W., Ben, Q., Wang, T., Mi, B., Huang, T., Zhao, S., Chen, Y., Yang, S., Cao, P., Yu, W., Ye, Z., Li, J., Long, J., Wang, Z., Wang, H., Zhao, Y., Tu, Z., Qiao, Y., Lin, D., Jiangmiao, P.: Grutopia: Dream general robots in a city at scale. In: arXiv (2024)
26. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
27. Wang, J., Chen, M., Zhang, S., Karaev, N., Schönberger, J., Labatut, P., Bojanowski, P., Novotny, D., Vedaldi, A., Rupprecht, C.: VGGT-Ω. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
28. Weng, Y., Wen, B., Tremblay, J., Blukis, V., Fox, D., Guibas, L., Birchfield, S.: Neural implicit representation for building digital twins of unknown articulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3141–3150 (2024)
29. Werby, A., Büchner, M., Röfer, A., Huang, C., Burgard, W., Valada, A.: Articulated object estimation in the wild. In: Conference on Robot Learning (CoRL). vol. 2 (2025)
30. Wu, X., Guo, Y., Chen, X., Yang, L., Lu, C., Li, Y.L.: Revisiting articulated parts perception in robot manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2026)
31. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., Yi, L., Chang, A.X., Guibas, L.J., Su, H.: SAPIEN: A simulated

part-based interactive environment. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
32. Zhou, M., Li, R., Lyu, X., Song, Z., Huang, Z., Zheng, C., Rupprecht, C., Vedaldi, A., Wu, S.: Articraft: An agentic system for scalable articulated 3d asset generation. arXiv preprint arXiv:2605.15187 (2026)