UniPart: Part-Level 3D Generation with Unified 3D Geom-Seg Latents Supplementary Material

Xufan He[1,2*] Yushuang Wu[2*] Xiaoyang Guo[2] Chongjie Ye[4,5] Jiaqing Zhou[2]
Tianlei Hu[3] Xiaoguang Han[5,4,6] Dong Du[1†]
[1]Nanjing University of Science and Technology [2]ByteDance Games [3]Zhejiang University
[4]FNii-Shenzhen [5]SSE, CUHKSZ [6]Guangdong Provincial Key Laboratory of Future Networks of Intelligence

1. Implementation Details

Network Architecture Our model is built upon the SOTA VectSet-based Hunyuan3D-2.1 [10]. For the global-level DiT, we adopt a Transformer architecture aligned with Hunyuan-DiT [5], enhanced with Mixture-of-Experts (MoE) layers to improve capacity and efficiency. Image conditions are encoded via DINOv2 [8] and injected through cross-attention. For the part-level DiT, we use the same architecture but condition on both the image and latent representations. The latent condition consists of segmented global latents, encoding global geometåry and the target part mask. These latent vectors are processed by a lightweight 8-layer attention block and fused via cross-attention. All conditions are randomly dropped with a 10% probability to enable classifier-free guidance.

Segmentation Encoder We adapt the structure of VectSet [11] encoder, append a part id embedding for each point after the position and normal. The part id embedding is drawn from a set of learnable embeddings, where the embedding corresponding to each part id is randomly selected, with the only constraint that different part ids correspond to distinct embedding vectors. Random selection avoids introducing spurious biases from predefined ID orderings, making the learned part representations more flexible and generalizable.

Segmentation Decoder We utilize a prompt-based segmentation mask decoder for generating an undetermined number of masks. Given a fixed number of latent tokens (e.g., 4096) and randomly sampled prompt tokens, a shallow MLP is adopted to predict whether each latent token belongs to the same part as the corresponding prompt token, thereby producing a segmentation mask. To achieve automatic segmentation, we perform random dense sampling on the latent tokens to generate sufficient prompt candidates, ensuring full coverage of all parts. After obtaining a large number of initial masks, we apply non-maximum suppression (NMS) as post-processing to remove duplicate and redundant masks, retaining only the most reliable predictions. In this way, our decoder can flexibly and automatically produce high-quality segmentation results without predefining the number of output masks.

Position Decoder For dense mask post-processing and visualization, we train a position decoder to regress the precise 3D coordinate corresponding to each latent token. The feasibility of this design stems from the inherent spatial locality of the VectSet latent tokens. Although VectSet does not impose explicit spatial locality constraints on the latent representations, we observe that each latent token naturally possesses localized characteristics owing to its generation mechanism. Specifically, it first samples dense point clouds, followed by furthest point sampling (FPS) to extract a fixed number of representative points (e.g., 4096). These FPS-sampled points act as the queries in cross-attention layers, while the original dense point clouds serve as keys and values, yielding fixed-length latent tokens. The tokens are then mapped to the latent space via several stacked self-attention layers. As each latent token is derived from the cross-attention query initialized by an individual FPS point, it inherently encodes localized geometric information. Based on this property, the position decoder can reliably reconstruct the corresponding spatial position for every latent token. This locality is also the prerequisite for us to perform segmentation on the latent tokens.

Dataset Curation We construct a curated dataset by integrating multiple public sources [2–4], yielding 300K objects with part-level segmentations. Duplicate vertices are merged using a tolerance of 10^{-6}, and meshes are split along connected components to isolate individual parts. Parts with fewer than 6 faces or total face area below 10^{-3} are removed to eliminate small, noisy floating artifacts, and only meshes with part counts between 2 and 32 are retained to ensure meaningful structural diversity. We filter out poorly segmented data, primarily from 3D scans, by manually reviewing explosion-view renderings, because connectivity-based segmentation is unreliable for scanned data. We further apply the winding number [1] for remesh-

ing, effectively achieving hole filling and gap closure to ensure that the mesh is watertight, and the part labels of the remeshed faces are determined based on the nearest face labels from the original raw mesh.

2. Comparison

Comparison with Closed-Source Commercial Models We additionally provide comparisons with closed-source commercial models. Some results are visualized in Fig. S3. As shown, although with a smaller-scale base model for geometry generation and much less training data for part segmentation, our UniPart can still achieve comparable part-level generation results.

Comparison on 3D Segmentation Our UniPart actually conducts part segmentation during generation, so the segmentation performance can not be directly evaluated on existing 3D part segmentation benchmarks. Therefore, we specifically construct an evaluation set with 50 object meshes generated by our whole-level DiT (decode the global-geometry latent with Geom. decoder). A group of ten human annotators is invited to manually annotate the part segmentation on these meshes, also given the input image as a reference. The collected annotations are further double-checked and finely corrected by annotators in the second round to guarantee high quality. As a result, we obtain 536 high-quality part annotations from the 50 object meshes. Based on them, we evaluate the part segmentation results of UniPart with existing SOTA baseline methods, including SAMesh [9], PartField [6], and P3-SAM [7]. For a direct comparison with ours, which only contains 4,096 latent vectors as part segmentation, we uniformly sample 4,096 points from each segmented point cloud or mesh surface of other methods to compute the numerical results. For quantitative results, we adopt mean IoU (mIoU) as the evaluation metric, with results shown in Tab. S1.

Table S1. Quantitative comparison on the geometry quality of part-level generation results. Best results are marked in bold font.

Method	SAMesh [9]	PartField [6]	P3-SAM [7]	**UniPart (Ours)**
mIoU ↑	0.3608	0.4167	0.7046	**0.7222**

3. Additional Results

We provide more generation results of UniPart, including global geometry, part latent segmentation, and part-level geometry, in Fig. S1 and Fig. S2 to show the impressive performance of our method and the robustness to various object images.

References

[1] Gavin Barill, Nia Dickson, Ryan Schmidt, David I.W. Levin, and Alec Jacobson. Fast winding numbers for soups and clouds. *ACM Transactions on Graphics*, 2018. 1

[2] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding, 2022. 1

[3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022.

[4] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 1

[5] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. 1

[6] Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. *arXiv preprint arXiv:2504.11451*, 2025. 2

[7] Changfeng Ma, Yang Li, Xinhao Yan, Jiachen Xu, Yunhan Yang, Chunshi Wang, Zibo Zhao, Yanwen Guo, Zhuo Chen, and Chunchao Guo. P3-sam: Native 3d part segmentation. *arXiv preprint arXiv:2509.06784*, 2025. 2

[8] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[9] George Tang, William Zhao, Logan Ford, David Benhaim, and Paul Zhang. Segment any mesh: Zero-shot mesh part segmentation via lifting segment anything 2 to 3d. *arXiv e-prints*, pages arXiv–2408, 2024. 2

[10] Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025. 1

[11] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *TOG*, 42(4):1–16, 2023. 1

Figure S1. Additional generation results using our UniPart. (a) Input images; (b) Whole object geometry generated by our whole-level DiT; (c) Assembled part meshes produced by our part-level DiT; (d) Visualization of part latent segmentation produced by our whole-level DiT.

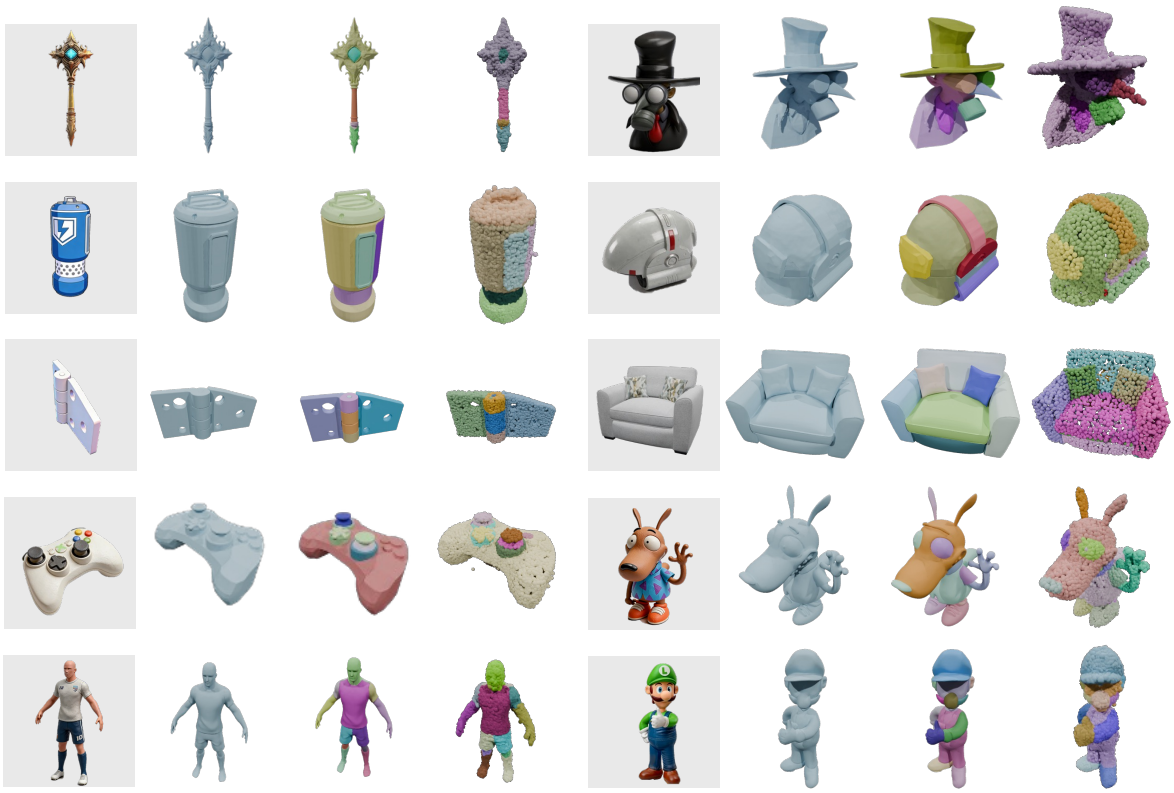

Figure S2. More generation results using our UniPart. (a) Input images; (b) Whole object geometry generated by our whole-level DiT; (c) Assembled part meshes produced by our part-level DiT; (d) Visualization of the part latent segmentation produced by our whole-level DiT.

Figure S3. Qualitative comparison with closed-source commercial models. We visualize the exploded views for better illustration of part-level generation. Our UniPart yields competitive results despite using a smaller-scale base model and less training data.