LaS-Comp: Zero-shot 3D Completion with Latent–Spatial Consistency
(Supplementary Material)

The supplementary material is organized as follows. We discuss the latent–spatial gap between partial shapes and ground truth in Sec. A, detail the evaluation metrics in Sec. B, and provide additional implementation details in Sec. C. Sec. D presents further details of the Omni-Comp benchmark. Additional qualitative visualizations are shown in Sec. E, and limitations and future directions are summarized in Sec. F.

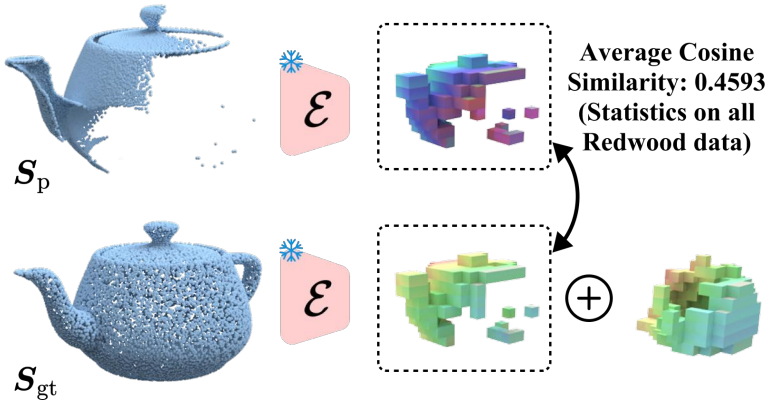

Figure A. Illustration of the latent gap between the partial input $\boldsymbol{S}_{\mathrm{p}}$ and the ground truth $\boldsymbol{S}_{\mathrm{gt}}$. Although $\boldsymbol{S}_{\mathrm{p}}$ and $\boldsymbol{S}_{\mathrm{gt}}$ share the same surface geometry in the observed regions, their latents obtained by the VAE encoder in the corresponding regions show an obvious difference.

A. More Discussions about the Latent-Spatial Gap

As mentioned in the main paper, we observe that the latents of the partial input and the ground truth obtained by the VAE encoder in the observed regions differ significantly. As shown in Fig. A, we obtain the latents of $\boldsymbol{S}_{\mathrm{p}}$ and $\boldsymbol{S}_{\mathrm{gt}}$ respectively. Then we utilize the latent mask $\boldsymbol{M}$ downsampled from the voxel of $\boldsymbol{S}_{\mathrm{p}}$ with resolution of 64^3 to the latent space with resolution of 16^3, to select the latents of both $\boldsymbol{S}_{\mathrm{p}}$ and $\boldsymbol{S}_{\mathrm{gt}}$ that belong to the partial regions, and compute the cosine similarity on the chosen latents. Computing such latent cosine similarity among all the samples with their GT on the Redwood dataset [5], we find that the average cosine similarity is only **0.4593**. As discussed in prior works [9, 15, 23], cosine similarities greater than 0.80 are generally regarded as reflecting strong correlations. In contrast, our measured average similarity of 0.4593 falls considerably below this threshold. This value indicates that the latent codes of the partial inputs and their corresponding GT shapes are far from strongly correlated in the overlapping regions.

This limited correlation suggests that directly relying on partial latent codes for completion guidance is fundamentally fragile, which is also verified by our ablation study in the main paper. Instead of performing completion in latent space, we project back to the original geometric space and explicitly inject the known partial points into the decoded space, so that the observed geometry is faithfully preserved and provides a stronger, more reliable constraint to guide the completion process.

B. Details of the Evaluation Metrics

Chamfer Distance (CD) quantitatively evaluates the quality of completed point clouds, by adopting the symmetric distance between the predicted completion $\boldsymbol{S}_{\mathrm{c}}$ and the ground-truth (GT) shape $\boldsymbol{S}_{\mathrm{gt}}$. Let $\boldsymbol{S}_{\mathrm{c}} = \{\boldsymbol{p}_i^{\mathrm{c}}\}_{i=1}^{N} \subset \mathbb{R}^3$ denote the completed

point cloud predicted by the model, and $\boldsymbol{S}_{\text{gt}} = \{\boldsymbol{p}_j^{\text{gt}}\}_{j=1}^{M} \subset \mathbb{R}^3$ denote the ground-truth point cloud. The Chamfer Distance is defined as the sum of two one-sided nearest-neighbor distances:

$$\text{CD}(\boldsymbol{S}_{\text{c}}, \boldsymbol{S}_{\text{gt}}) = \frac{1}{N}\sum_{i=1}^{N} \min_{1\le j\le M} \left\|\boldsymbol{p}_i^{\text{c}} - \boldsymbol{p}_j^{\text{gt}}\right\|_2 + \frac{1}{M}\sum_{j=1}^{M} \min_{1\le i\le N} \left\|\boldsymbol{p}_j^{\text{gt}} - \boldsymbol{p}_i^{\text{c}}\right\|_2. \tag{1}$$

Here $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^3$. The first term measures how well every predicted point is supported by the ground truth, while the second term measures how well the prediction covers the ground-truth surface. We report the metric in our experiments by $\times 10^2$.

Earth Mover's Distance (EMD) [17] is defined as the minimum average cost of transporting one set to the other. After uniformly resampling of $\boldsymbol{S}_{\text{c}}$ and $\boldsymbol{S}_{\text{gt}}$, we assume $|\boldsymbol{S}_{\text{c}}| = |\boldsymbol{S}_{\text{gt}}| = N$. The Earth Mover's Distance is defined as the minimum average cost of transporting one set to the other:

$$\text{EMD}(\boldsymbol{S}_{\text{c}}, \boldsymbol{S}_{\text{gt}}) = \min_{\phi:\, \boldsymbol{S}_{\text{c}} \to \boldsymbol{S}_{\text{gt}}} \frac{1}{|\boldsymbol{S}_{\text{c}}|} \sum_{\boldsymbol{p}\in\boldsymbol{S}_{\text{c}}} \left\|\boldsymbol{p} - \phi(\boldsymbol{p})\right\|_2, \tag{2}$$

where ϕ is a bijection between $\boldsymbol{S}_{\text{c}}$ and $\boldsymbol{S}_{\text{gt}}$. This bijective matching is indicative of both the uniformity and the geometric fidelity of the generated shapes. Following ComPC [11], we set eps as 0.005, iteration as 50 for the computation of EMD. We report the metric in our experiments by $\times 10^2$.

Unidirectional Chamfer Distance (UCD) measures the squared L_2 distance from the partial input shape $\boldsymbol{S}_{\text{p}}$ to the completed output $\boldsymbol{S}_{\text{c}}$. Given the partial point set $\boldsymbol{S}_{\text{p}} \subset \mathbb{R}^3$ and the completed point set $\boldsymbol{S}_{\text{c}} \subset \mathbb{R}^3$, we define

$$\text{UCD}(\boldsymbol{S}_{\text{p}}, \boldsymbol{S}_{\text{c}}) = \frac{1}{|\boldsymbol{S}_{\text{p}}|} \sum_{\boldsymbol{p}\in\boldsymbol{S}_{\text{p}}} \min_{\boldsymbol{q}\in\boldsymbol{S}_{\text{c}}} \left\|\boldsymbol{p} - \boldsymbol{q}\right\|_2^2. \tag{3}$$

Unlike the symmetric Chamfer Distance, UCD only measures how well the completed shape preserves and explains the observed partial input. We report the metric in our experiments by $\times 10^4$.

Unidirectional Hausdorff Distance (UHD) [4, 20] similarly measures the single-sided Hausdorff distance from the partial input shape $\boldsymbol{S}_{\text{p}}$ to the completed output $\boldsymbol{S}_{\text{c}}$:

$$\text{UHD}(\boldsymbol{S}_{\text{p}}, \boldsymbol{S}_{\text{c}}) = \max_{\boldsymbol{p}\in\boldsymbol{S}_{\text{p}}} \min_{\boldsymbol{q}\in\boldsymbol{S}_{\text{c}}} \left\|\boldsymbol{p} - \boldsymbol{q}\right\|_2. \tag{4}$$

Compared with UCD, UHD focuses on the worst-case (hardest-to-match) observed point in $\boldsymbol{S}_{\text{p}}$. We report the metric in our experiments by $\times 10^2$.

Minimum Matching Distance (MMD) [1] measures the fidelity of a set of generated completed shapes with respect to the set of ground-truth shapes. Let $\mathcal{D}_{\text{c}} = \{\boldsymbol{S}_{\text{c}}^{(i)}\}_{i=1}^{N_{\text{c}}}$ be the set of completed shapes and $\mathcal{D}_{\text{gt}} = \{\boldsymbol{S}_{\text{gt}}^{(j)}\}_{j=1}^{N_{\text{gt}}}$ be the set of ground-truth shapes, where each $\boldsymbol{S}_{\text{c}}^{(i)}$ and $\boldsymbol{S}_{\text{gt}}^{(j)}$ is a point cloud in $\mathbb{R}^3$. Given a point-set distance $\text{CD}(\cdot,\cdot)$, the Minimum Matching Distance is defined as

$$\text{MMD}(\mathcal{D}_{\text{c}}, \mathcal{D}_{\text{gt}}) = \frac{1}{|\mathcal{D}_{\text{gt}}|} \sum_{\boldsymbol{S}_{\text{gt}}\in\mathcal{D}_{\text{gt}}} \min_{\boldsymbol{S}_{\text{c}}\in\mathcal{D}_{\text{c}}} \text{CD}\big(\boldsymbol{S}_{\text{gt}}, \boldsymbol{S}_{\text{c}}\big). \tag{5}$$

That is, for each ground-truth shape we find the closest generated completion in $\mathcal{D}_{\text{c}}$ under $\text{CD}(\cdot,\cdot)$ and report the average matched distance. We report the metric in our experiments by $\times 10^2$.

Total Mutual Difference (TMD) measures the diversity of multiple completion results generated for the same partial input. Given a partial shape $\boldsymbol{S}_{\text{p}}$, we generate a set of K completed shapes $\mathcal{D}_{\text{c}} = \{\boldsymbol{S}_{\text{c}}^{(k)}\}_{k=1}^{K}$, where each $\boldsymbol{S}_{\text{c}}^{(k)}$ is a point cloud in $\mathbb{R}^3$. We apply the Chamfer Distance defined in Eq. (1) to assess the difference of each pair of point clouds inside $\mathcal{D}_{\text{c}}$. The Total Mutual Difference for this sample is defined as the average pairwise distance among all completions:

$$\text{TMD}(\mathcal{D}_{\text{c}}) = \frac{2}{K(K-1)} \sum_{1\le a<b\le K} \text{CD}\big(\boldsymbol{S}_{\text{c}}^{(a)}, \boldsymbol{S}_{\text{c}}^{(b)}\big). \tag{6}$$

A larger TMD value indicates higher diversity among the generated completion hypotheses for the same partial input $\boldsymbol{S}_{\text{p}}$. We report the metric in our experiments by $\times 10^2$.

C. More Implementation Details

C1. Backbones for 3D Completion

Our framework is compatible with backbones that expose three abstract components: a spatial-to-latent encoder ($\mathcal{E}$), a latent-to-spatial decoder ($\mathcal{D}$), and a latent generative process ($\mathcal{G}$). We instantiate this with Direct3D-S2 [21] and TRELLIS [22]. For Direct3D-S2, which operates on the dense voxel grid, $\mathcal{E}$ is its VAE encoder, $\mathcal{G}$ is the conditional diffusion transformer, and $\mathcal{D}$ is the VAE decoder. For TRELLIS, which uses models for the first stage (sparse structure generation), $\mathcal{E}$ is the VAE Encoder, $\mathcal{G}$ is the first stage of its rectified-flow transformer generator, and $\mathcal{D}$ is the VAE Decoder. For both backbones, the resolution of the voxel grid is 64^3; all the input to the models should be voxelized first. Our pipeline operates on the first stage of both backbones, which operates to generate dense voxel grids as the main geometry, while freezing the second stage of the models to generate SDF for meshes and sampling as point clouds for evaluation.

C2. Pipeline Parameter Settings

In our experiments, we set the number of denoising time steps to 100, the CFG scale to 1.0, and the rescale factor of t to 3.0. At each time step, we perform one-step latent optimization update with a learning rate of 1×10^{-5}. The occupancy threshold is set to 0.5, following the backbone settings [21, 22]. In the IAS stage, the output of the decoder $\mathcal{D}$ is kept as logits rather than thresholded into an occupancy grid; we directly compute the binary cross-entropy loss on these logits and optimize the latent. Our PNS with new sampling noise is applied only for $t \in [0.5, 1]$, while ERS and IAS are applied for all time steps.

C3. Dataset Parameter Settings

Following prior experimental settings [4, 6, 11, 24], we set the point cloud resolution of the Redwood [5] and synthetic [11, 13] datasets to 16,384, and that of KITTI [10] and ScanNet [7] to 2,048. Our proposed Omni-Comp benchmark also uses a resolution of 16,384.

C4. Work on consumer-grade GPUs.

Our method can be implemented on a single RTX3090: under FP16 with batch size 1, it uses 10.29 GB peak VRAM and runs in 35.92s, which demonstrates the practical deployability.

D. Details of the Omni-Comp Benchmark

We introduce **Omni-Comp**, a new benchmark designed for a more comprehensive and robust evaluation of 3D shape completion. Our benchmark features a challenging set of 30 objects, each from a distinct category (listed in Tab. A), curated from diverse sources: 10 real-world scans from Redwood [5] (chosen for complex geometry), 10 real-world everyday objects from YCB [3] (motivated by downstream applications, *e.g.* , robotic grasping), and 10 synthetic shapes from [18] (chosen for rich semantic structure, with complex geometry). Critically, inspired by [24], our benchmark generates three distinct partial patterns for each object: (i) *Single Scan*: using the projection of the captured depth map for real-world data, and simulating a standard depth camera capture for synthetic data; (ii) *Random Crop*: Representing arbitrary occlusions by randomly cropping a portion; and (iii) *Semantic Part*: Keeping a semantic component and removing other parts. By creating two samples for each pattern per object, the benchmark comprises 180 challenging partial samples with corresponding ground truths.

When selecting real-world data from [3, 5], we intentionally avoid near-cuboidal or purely box-like geometries and instead prioritize objects with richer, more complex structures. For the synthetic data, we choose samples whose identities do not appear in the released Objaverse-XL [8] metadata or index, so that they are not included in the training data of our generative backbones.

We introduce the preprocessing of samples for the three partial patterns as follows. **(a)** When constructing samples of a single scan, if not providing the object mask, we utilize SAM2 [16] and OWL-ViT [14] to extract the mask in the real-world RGB-D data, where the depth map is registered to the image. After that, we back-project the depth according to the mask to the 3D space. Next, we follow SDS-Comp [12], manually align the provided ground truth point clouds with the back-projected partial point cloud, and apply ICP for refinement [2]. The ground truth with the aligned partial scan is normalized by the bounding box of GT, to the range of $[-0.5, 0.5]$. For the synthetic data, we establish a virtual camera with virtual scans to construct the partial point clouds. **(b)** Considering the random crop, we sample axis-aligned slabs and half-space cuts along random axes. Concretely, we select percentile-based windows or half-spaces along a chosen axis to retain only a random fraction of the GT object (e.g., 40-70%), so that each crop mimics realistic partial views where only a contiguous portion of the object is observed. **(c)** Speaking of the semantic part pattern, we manually isolate meaningful object parts on the complete object and remove all remaining geometry using an interactive editing tool. This yields semantic partial shapes

Table A. **Category list in the Omni-Comp benchmark.** Representative object categories from Redwood [5], YCB [3], and Synthetic [18].

Redwood		*YCB*		*Synthetic*	
ID	**Category**	**ID**	**Category**	**ID**	**Category**
1	Bicycle	11	Cracker Box	21	Dinosaur
2	Plant	12	Mustard Bottle	22	Glass
3	Trash Bin	13	Banana	23	Guitar
4	Car	14	Bleach Cleanser	24	Headphone
5	Motorcycle	15	Skillet	25	Laptop
6	Sign	16	Power Drill	26	Robot
7	Bench	17	Extra Large Clamp	27	Shoe
8	Couch	18	Tennis Ball	28	Statue
9	Desk	19	Wood Block	29	Toilet
10	Rocking Chair	20	Timer	30	Torch

that consist of a single, coherent part, providing a distinct partial pattern compared with the single scan and random crop patterns.

Note that all the partial and GT data are sampled with a resolution of 16,384, following [11, 12]. We provide more visualizations of our dataset across diverse objects and different partial patterns, as shown in Fig. C and Fig. D.

E. More Visualization Results

E1. Visualization Results on the Redwood Dataset

To ensure a better understanding of our method's superior performance and a fair comparison setting, we provide more examples from the Redwood dataset [5], visualized as point clouds, in Fig. E. These real-world scans are highly challenging, containing strong sensor noise, self-occlusions, and depth discontinuities. Across a wide range of object instances and viewpoints, our method consistently recovers geometrically detailed and semantically plausible completions: the global structure of the object is preserved, while fine-scale parts such as legs, handles, and support structures are reasonably completed.

E2. Visualization Results on the Synthetic Dataset

We also provide qualitative examples on synthetic data from [11], as visualized in Fig. F. Compared with real-world scans, these synthetic partial point clouds exhibit cleaner sampling and more diverse geometric patterns, including thin structures, high-curvature regions, and complex topologies. Our method is able to accurately complete these partial observations into full shapes with rich geometric details and coherent semantics, showing that the proposed pipeline generalizes well to different domains of data.

E3. Visualization Results on the Proposed Omni-Comp Benchmark

To further demonstrate our method's generalization ability across diverse partial patterns and object categories, we provide additional qualitative examples on our Omni-Comp benchmark in Fig. G, Fig. H, and Fig. I, respectively. The benchmark covers multiple partial patterns (random crop, single scan, and semantic part) and spans a wide spectrum of categories. Under all these settings, our method produces geometrically detailed and semantically meaningful completions that respect the observed regions while plausibly hallucinating the missing parts. These results highlight the strong robustness and cross-pattern generalization of our framework when faced with heterogeneous partial inputs.

E4. Visualization Results on Completion Diversity

Finally, to verify that our method can generate diverse yet reasonable completions for the same partial input, we present additional qualitative examples in Fig. J under both unconditional completion and text-guided completion. Given a fixed partial point cloud, our model is able to produce multiple distinct full-shape hypotheses that all remain consistent with the observed geometry, illustrating its ability to capture the inherent ambiguity of the completion task. These visualizations demonstrate that our approach not only yields high-quality and geometrically refined completions but also supports meaningful multi-modal diversity in the output space.

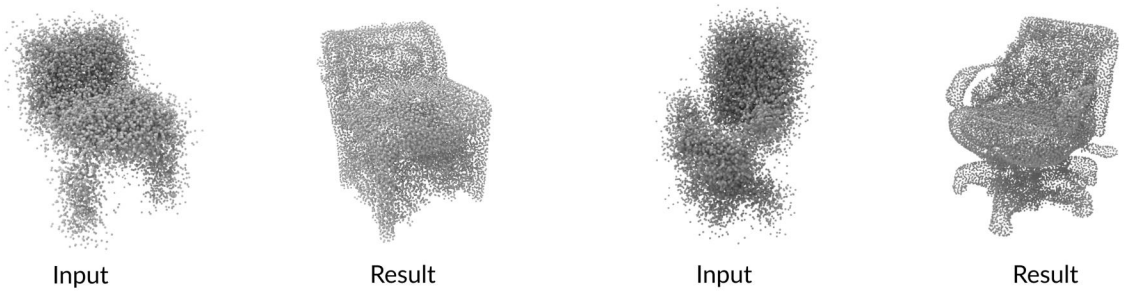

Figure B. Visual examples of the completion results under extremely noisy partial inputs. Despite the severe noise that heavily corrupts the observed points, our method can still recover a reasonable global structure and overall object silhouette, but many fine details and thin structures are degraded or missing, revealing the limitation of our model when strong noise overwhelms the underlying geometry.

F. Limitations and Future Work

Although our method outperforms existing approaches in various benchmarks, object categories, and partial patterns, extremely noisy inputs remain challenging and may still lead to imperfect completions; see Fig. B. In such cases, the model can still recover the coarse object structure, *e.g.* , overall chair silhouette, but fine details and thin structures may be over-smoothed or distorted because the underlying geometry is heavily corrupted. When the input contains only weak or ambiguous cues, the generative prior has limited reliable information to condition on. To further improve robustness in such scenarios, our future work will explore: a) stronger outlier-removal designs that explicitly detect and remove noise patterns such as scattered points far from the main shape or clusters with inconsistent local geometry before completion, thereby exposing a cleaner partial input for shape completion; and b) confidence-aware refinement strategies that adapt the denoising strength based on the model's estimated reliability in each region, preserving confident areas while applying more cautious updates to ambiguous or noisy regions.

What's more, following ComPC, SDS-Comp, and GenPC, we adopt GT-based normalization. However, for real-world application and implementation, we usually encounter situations where only partial scans can be obtained, while the gt shapes are missing. In the future, we plan to develop more robust methods for such in-the-wild 3D completion.

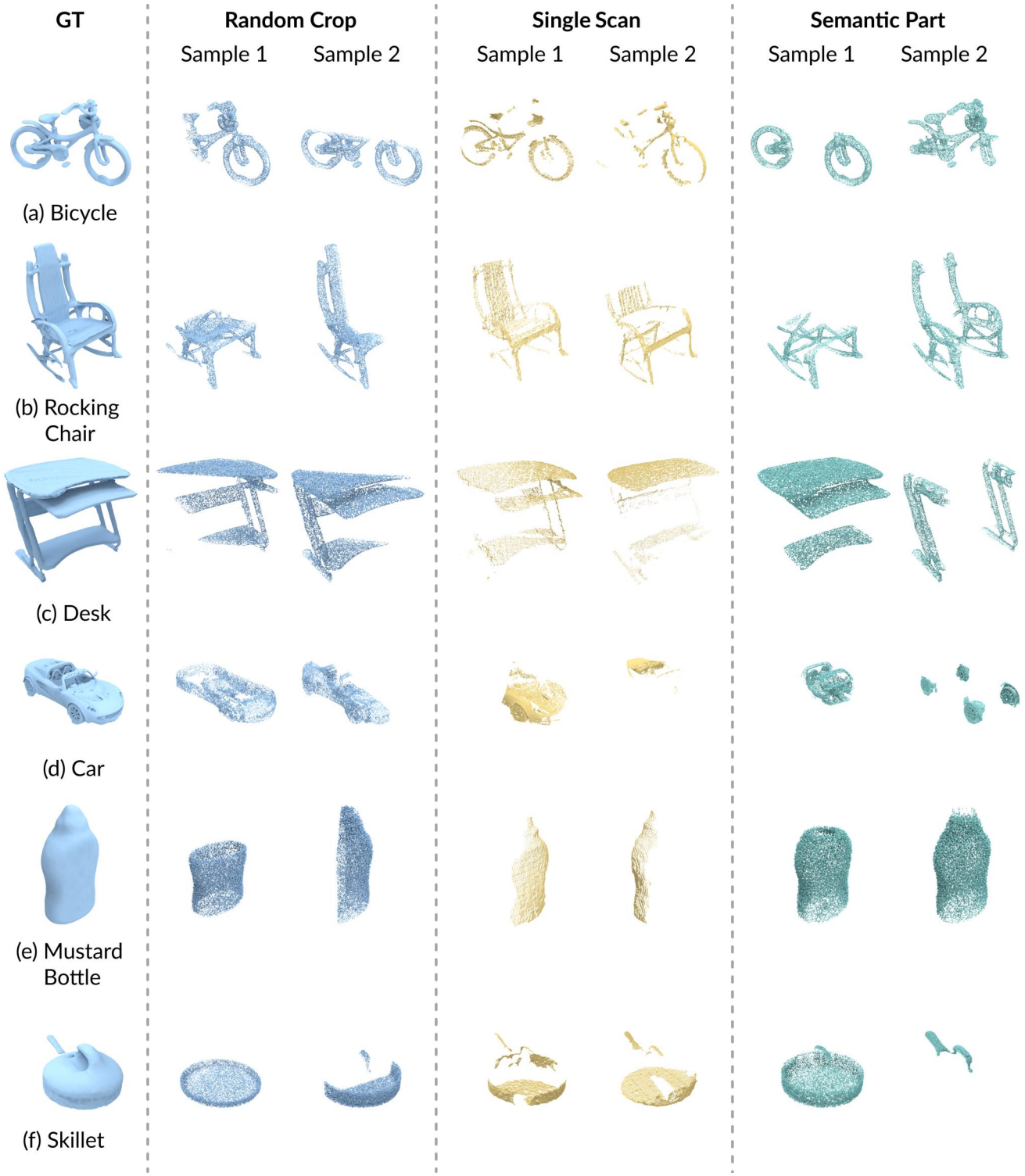

Figure C. Visual examples of the proposed Omni-Comp benchmark. We show the ground truth mesh, with its corresponding partial samples of random crop, single scan, and semantic part.

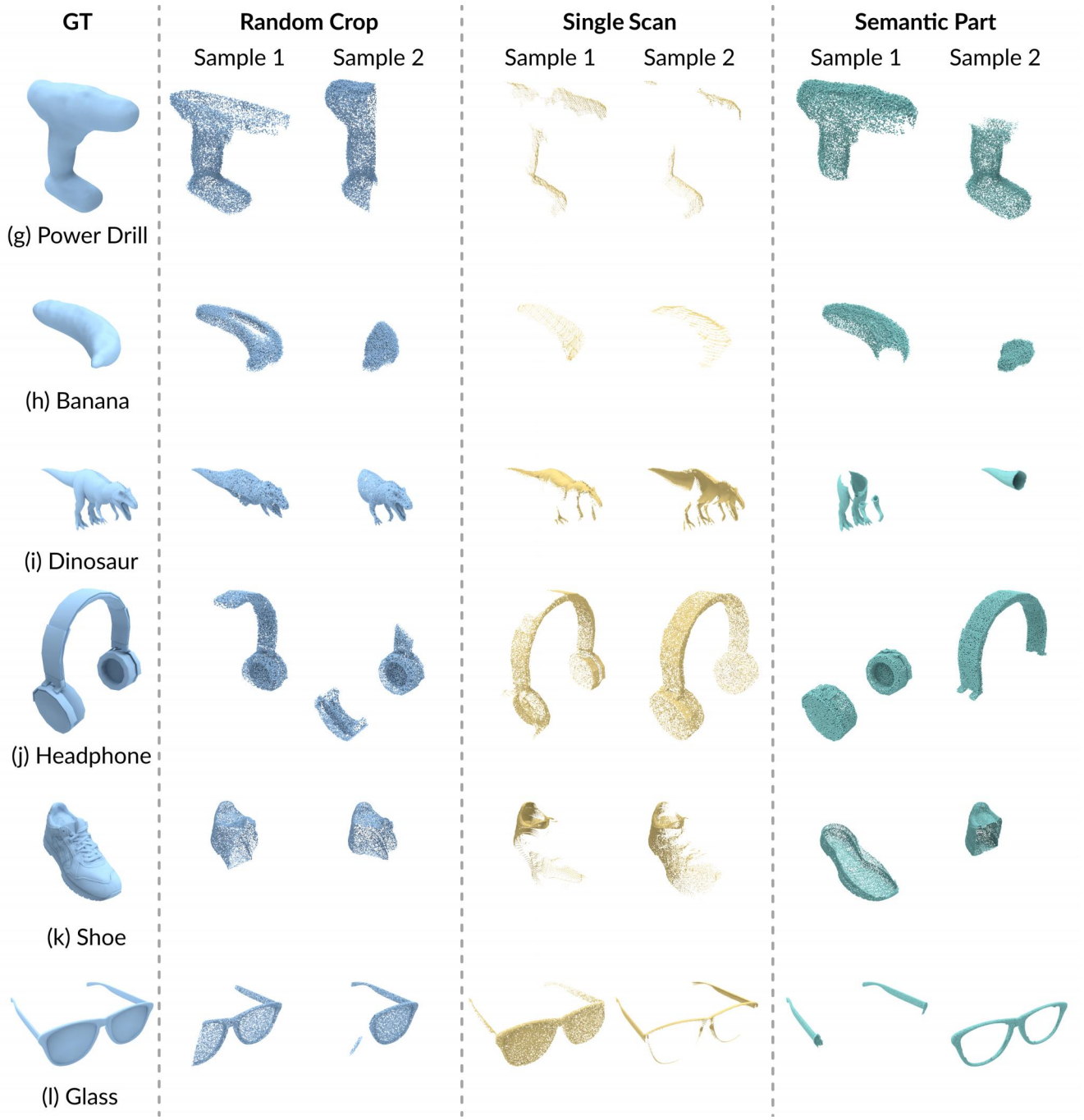

Figure D. Visual examples of the proposed Omni-Comp benchmark. We show the ground truth mesh, with its corresponding partial samples of random crop, single scan, and semantic part.

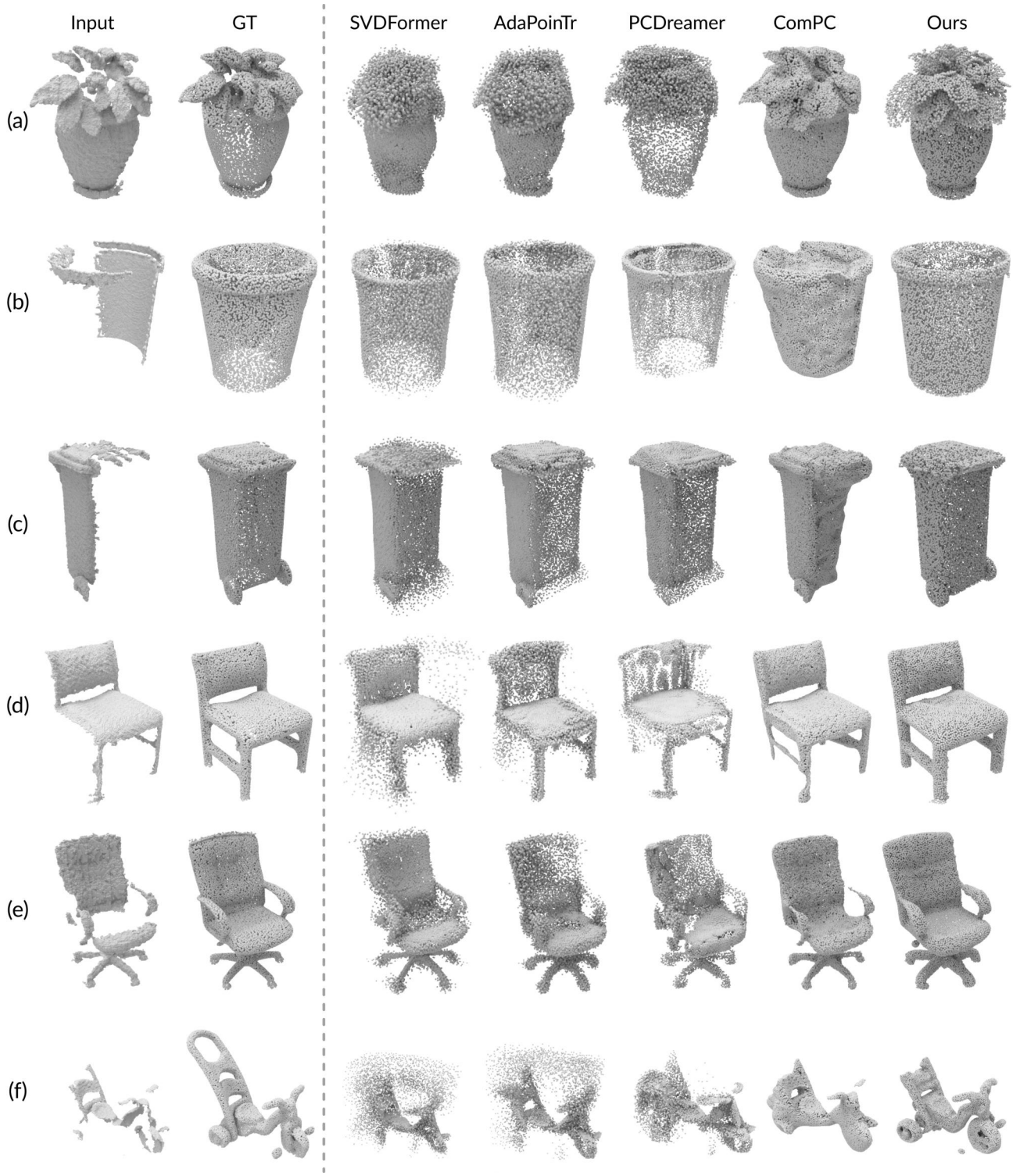

Figure E. Visual Comparison on the Redwood dataset [5], with point cloud representation. Obviously, our method significantly outperforms prior approaches in most of the examples, with very high-quality geometry and good completion correctness.

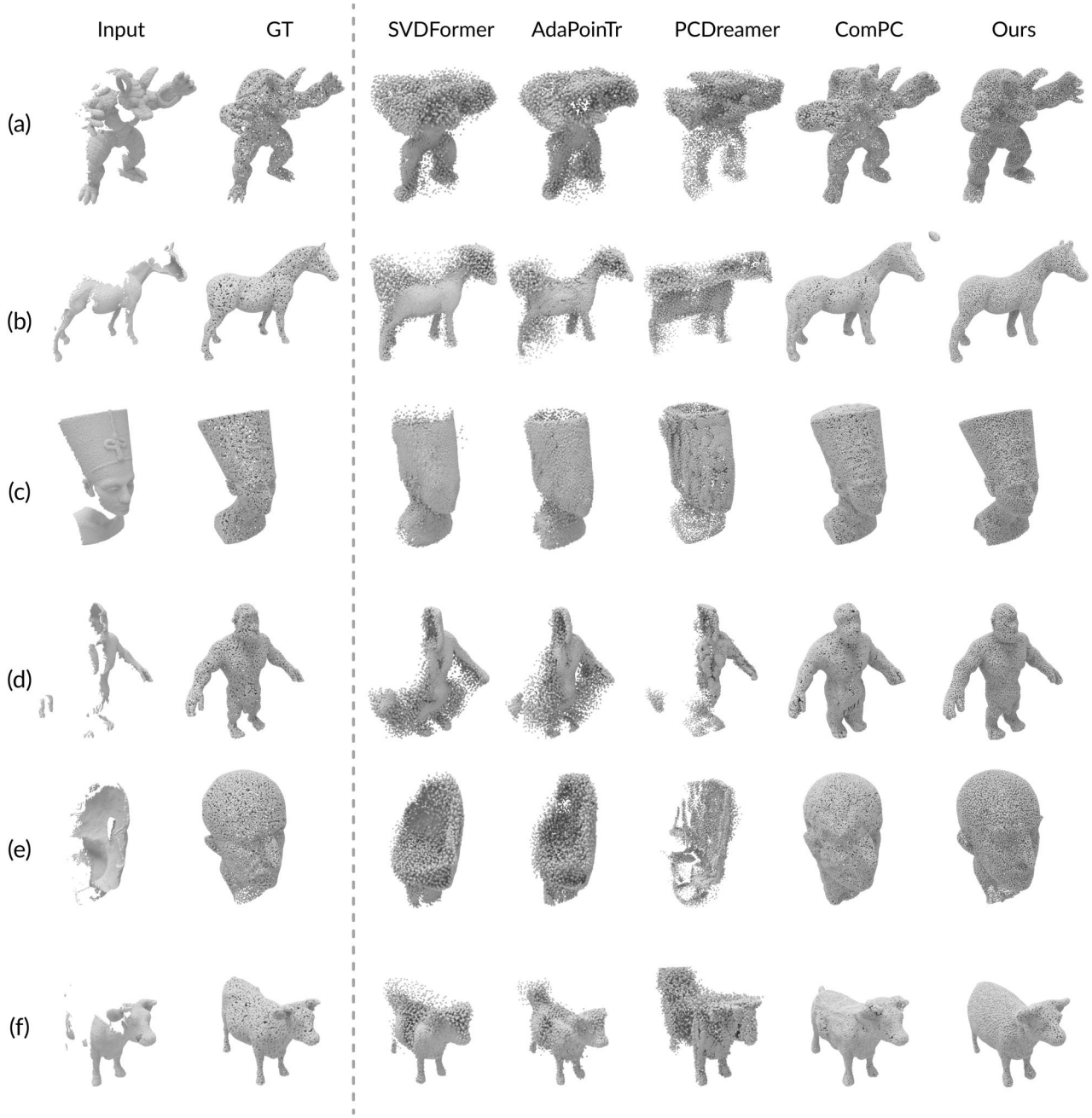

Figure F. Visual Comparison on the synthetic dataset [11], with point cloud representation. Obviously, our method significantly outperforms prior approaches in most of the examples, with very high-quality geometry and good completion correctness.

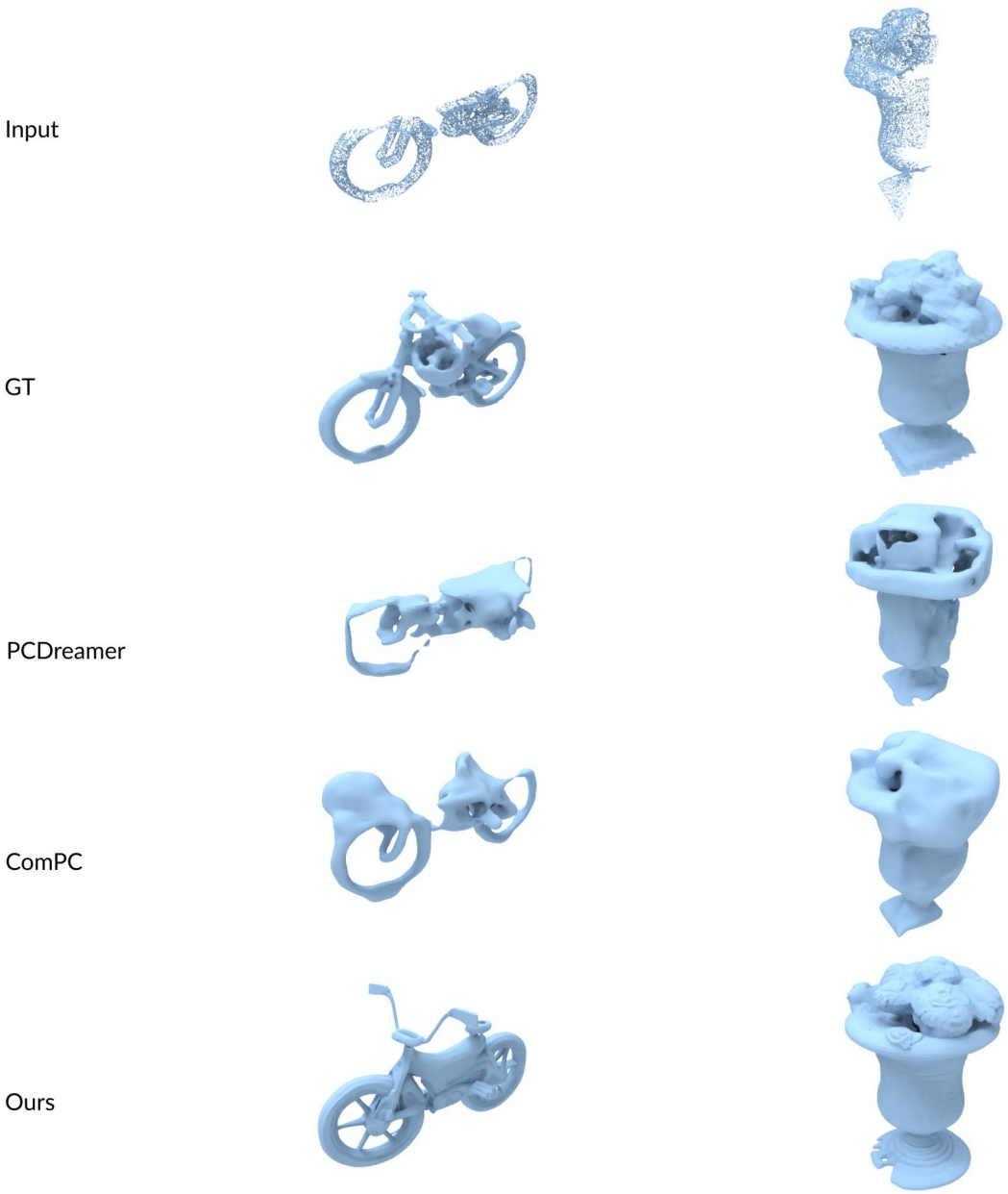

Figure G. More visual examples on the proposed Omni-Comp benchmark, under the random crop partial pattern. Our method consistently provides better completion results compared with the latest methods [11, 19], regardless of the partial pattern and object category.

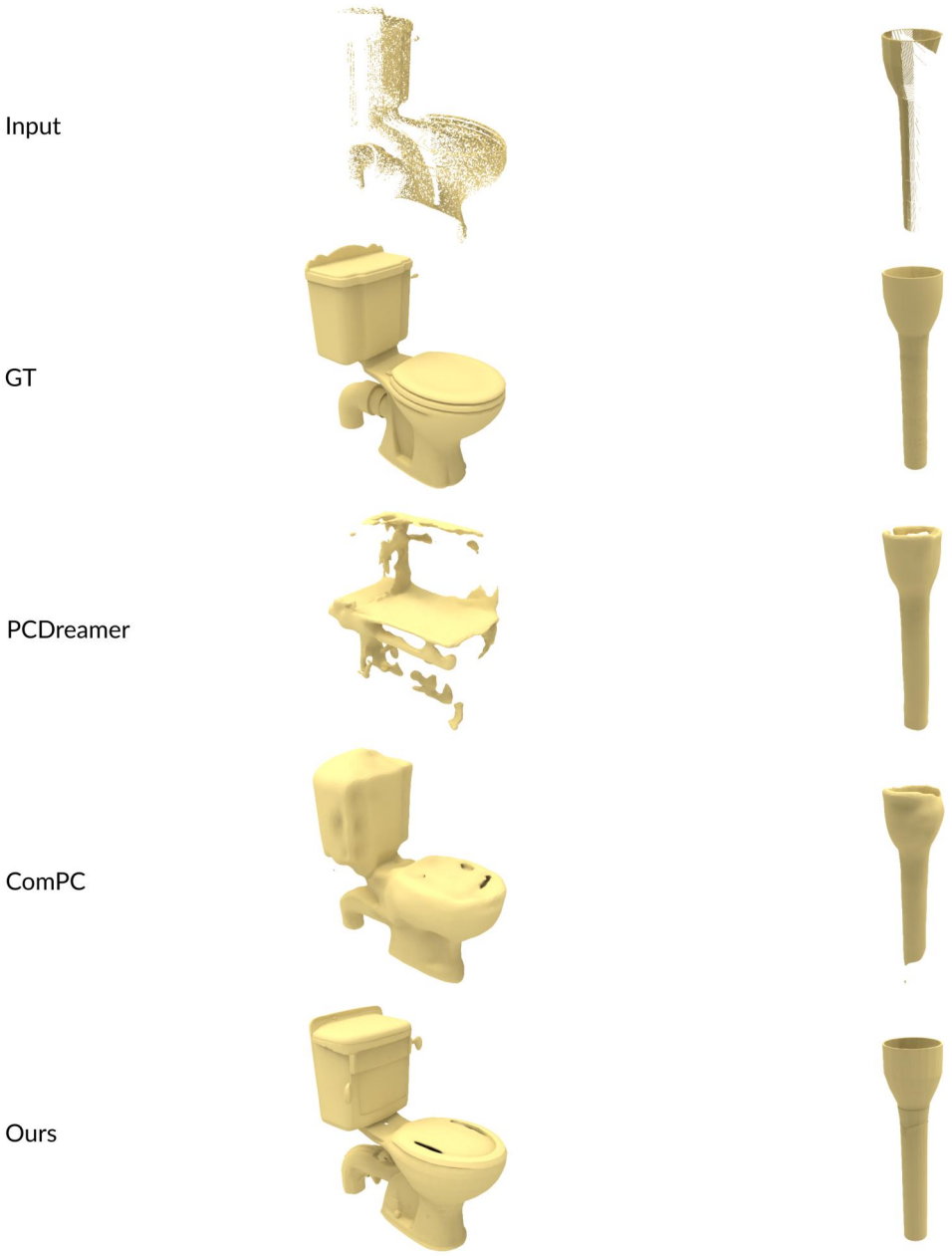

Figure H. More visual examples on the proposed Omni-Comp benchmark, under the single scan partial pattern. Our method consistently provides better completion results compared with the latest methods [11, 19], regardless of the partial pattern and object category.

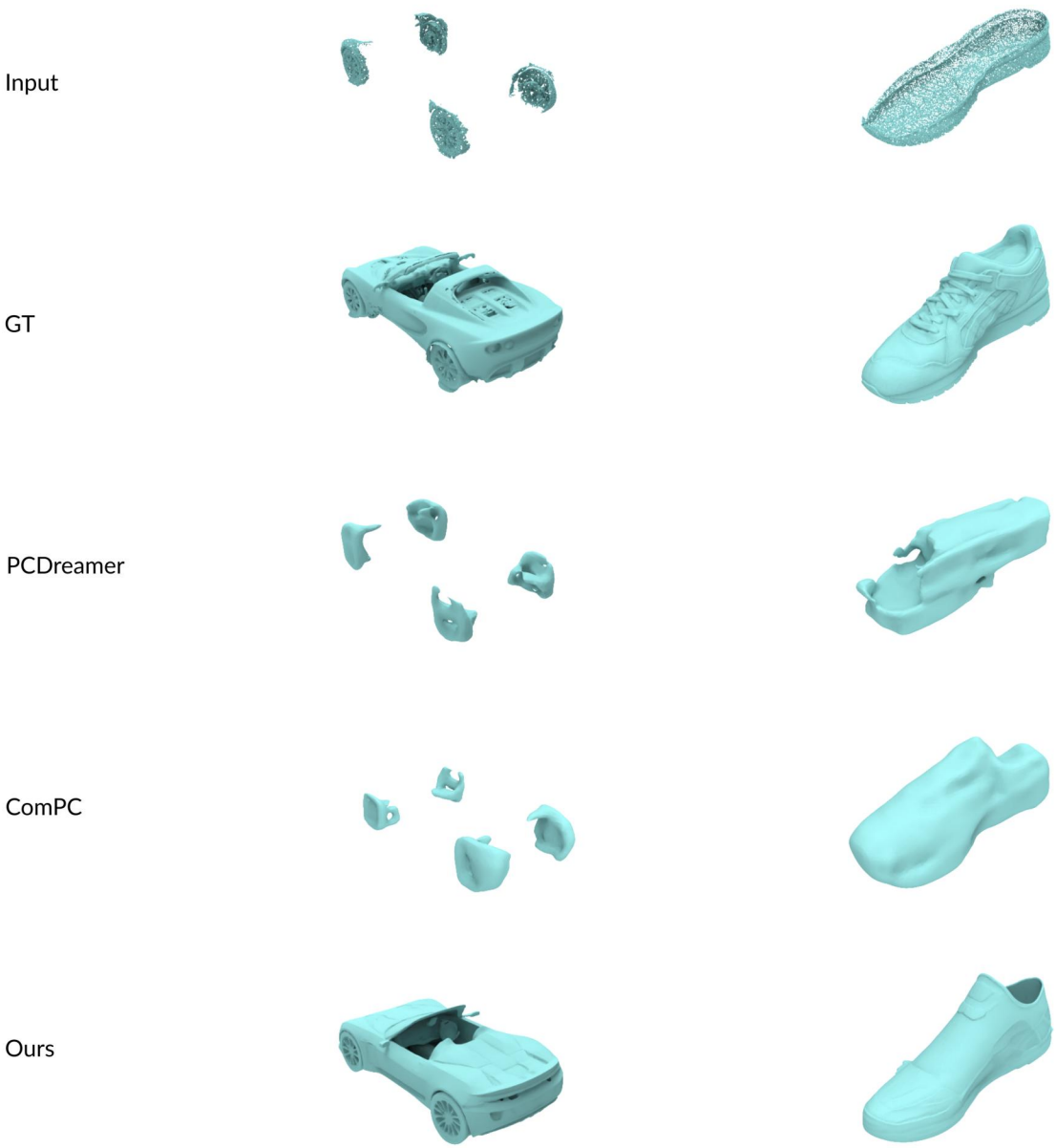

Figure I. More visual examples on the proposed Omni-Comp benchmark, under the semantic part partial pattern. Our method consistently provides better completion results compared with the latest methods [11, 19], regardless of the partial pattern and object category.

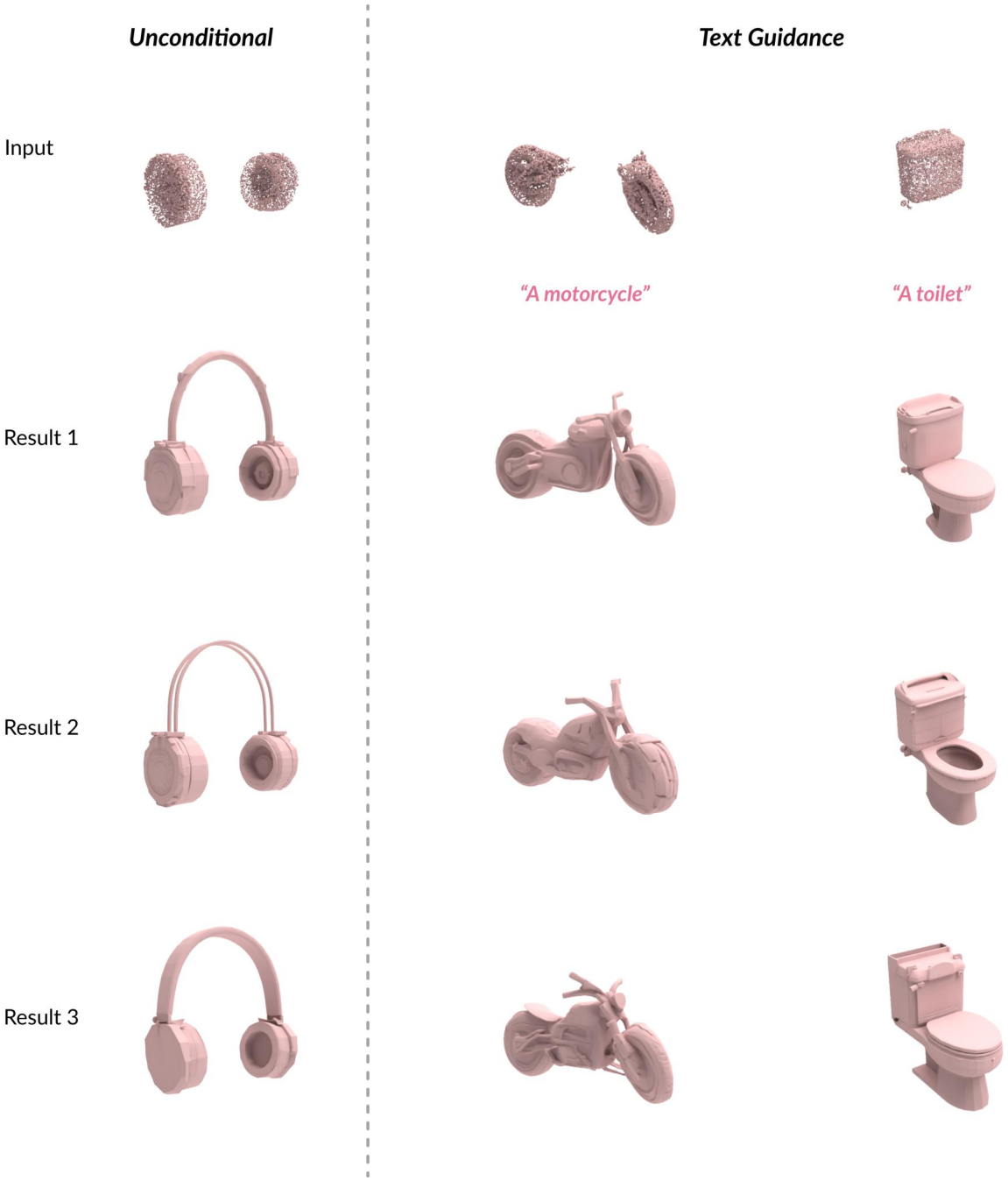

Figure J. More visual examples to show the completion diversity, under unconditional and text-guided completion. In both completion settings, our method can provide reasonable results with good geometry.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J. Guibas. Learning Representations and Generative Models for 3D Point Clouds. In *Proceedings of International Conference on Machine Learning (ICML)*, 2017. 2

[2] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-Squares Fitting of Two 3-D Point Sets. *IEEE Transactions Pattern Analysis & Machine Intelligence*, PAMI-9(5):698–700, 1987. 3

[3] Berk Çalli, Aaron Walsman, Arjun Singh, Siddhartha S. Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols. In *International Conference on Advanced Robotics (ICAR)*, 2015. 3, 4

[4] Xuelin Chen, Baoquan Chen, and Niloy J Mitra. Unpaired Point Cloud Completion on Real Scans using Adversarial Training. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3

[5] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A Large Dataset of Object Scans. *arXiv:1602.02481*, 2016. 1, 3, 4, 8

[6] Ruikai Cui, Shi Qiu, Saeed Anwar, Jiawei Liu, Chaoyue Xing, Jing Zhang, and Nick Barnes. P2C: Self-supervised Point Cloud Completion from Single Partial Clouds. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14351–14360, 2023. 3

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017. 3

[8] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663*, 2023. 3

[9] James D. Evans. *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole, Pacific Grove, CA, 1996. 1

[10] Andreas Geiger. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3354–3361, 2012. 3

[11] Tianxin Huang, Zhiwen Yan, Yuyang Zhao, and Gim H Lee. ComPC: Completing a 3D Point Cloud with 2D Diffusion Priors. In *International Conference on Learning Representations (ICLR)*, pages 51765–51784, 2025. 2, 3, 4, 9, 10, 11, 12

[12] Yoni Kasten, Ohad Rahamim, and Gal Chechik. Point Cloud Completion with Pretrained Text-to-image Diffusion Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 12171–12191, 2023. 3, 4

[13] Yaron Lipman, David Levin, and Daniel Cohen-Or. Green Coordinates. *ACM Transactions on Graphics*, 27(3):1–10, 2008. 3

[14] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[15] John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. Reevaluating Adversarial Examples in Natural Language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, 2020. 1

[16] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[17] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal Computer Vision*, 2000. 2

[18] Penghao Wang, Yiyang He, Xin Lv, Yukai Zhou, Lan Xu, Jingyi Yu, and Jiayuan Gu. PartNeXt: A Next-Generation Dataset for Fine-Grained and Hierarchical 3D Part Understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 3, 4

[19] Guangshun Wei, Yuan Feng, Long Ma, Chen Wang, Yuanfeng Zhou, and Changjian Li. PCDreamer: Point Cloud Completion Through Multi-view Diffusion Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27243–27253, 2025. 10, 11, 12

[20] Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal Shape Completion via Conditional Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV)*, pages 281–296, 2020. 2

[21] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Philip Torr, Xun Cao, and Yao Yao. Direct3D-S2: Gigascale 3D Generation Made Easy with Spatial Sparse Attention. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 3

[22] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21469–21480, 2025. 3

[23] Jin Yong Yoo and Yanjun Qi. Towards Improving Adversarial Training of NLP Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956, 2021. 1

[24] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3D Shape Completion through GAN Inversion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3