Order Matters: 3D Shape Generation from Sequential VR Sketches

Supplementary Material

In this appendix, we further justify the design choice and evaluate the generalization capabilities of our model. First, we provide additional ablation studies on Fourier features and sketch orders (Sec. A-1), and an experiment on shape completion (Sec. A-2). Next, we test sketches drawn without our surface-snapping tool (Sec. A-3), followed by free-hand sketches created without any reference shape (Sec. A-4). We then analyze the impact of the sketcher expertise (Sec. A-6) and the performance–speed tradeoff of our architecture (Sec. A-7). Finally, we provide additional qualitative illustrations of shape generation and sketch completion for both our method and competing baselines.

A-1. Additional Ablation

We further test the effectiveness of proposed 3D fourier features. As shown in Tab. A-1[A], replacing 3D spatial Fourier features with raw coordinates leads to a clear performance drop, confirming their importance for encoding fine geometric detail. Replacing 1D Fourier encodings with fixed positional embeddings also degrades performance and restricts variable-length sequences.

We also test different ordering strategy. Needless to say, our synthetic sketches are not intended to model human drawing behavior, but to provide effective supervision for learning a sketch-to-shape mapping. Stroke order is therefore treated as an inductive bias, not as a claim about how humans draw. For instance, DFS ordering empirically outperforms BFS (Tab. A-1[B]), but this does not necessarily reflect human sketching strategies.

We further evaluate the impact of order perturbations (Tab. A-1[C]): reversing stroke order has little effect, scrambling strokes causes a moderate drop, while scrambling points leads to a clear degradation in F-score. These results indicate that *consistent sequential structure*, rather than a specific human-like ordering, is what benefits learning.

Importantly, order modeling is critical for partial sketches: when reconstructing shapes from only the first half of a human-drawn sketch, our order-aware model outperforms an order-agnostic variant by +6.6 F1-score. We will therefore reframe our claim of "order matters" to these empirically supported points.

A-2. Cross-Modal Shape Completion

We evaluate our model's ability to infer complete 3D shapes from partial sketches. To simulate incomplete inputs during inference, we keep only the first fraction of points in the sketch sequence, preserving its natural drawing order, and pad the remainder with learned `MASK` tokens. `SEP` tokens are randomly inserted to mimic the natural stroke-length distribution of real sketches. The model then predicts the full 3D shape directly from these partially masked sequences.

Table A-1. **Additional Ablation.**

	Experiment	F-score ↑	CD×1000 ↓
	Best	**56.8**	**5.1**
A	No 3D Fourrier	52.1	5.6
A	No 1D Fourrier	48.2	6.3
B	BFS traversal	47.8	6.4
C	Reverse Order	56.2	**5.0**
C	Scrambled Strokes	54.6	5.9
C	Scrambled Points	52.2	5.2

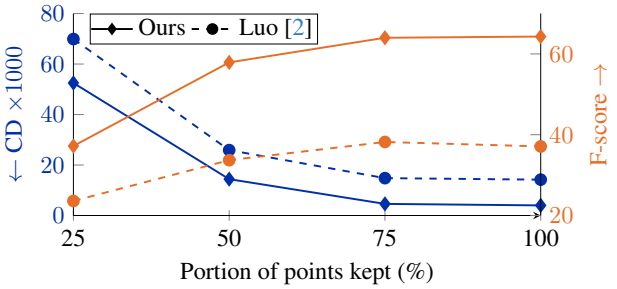


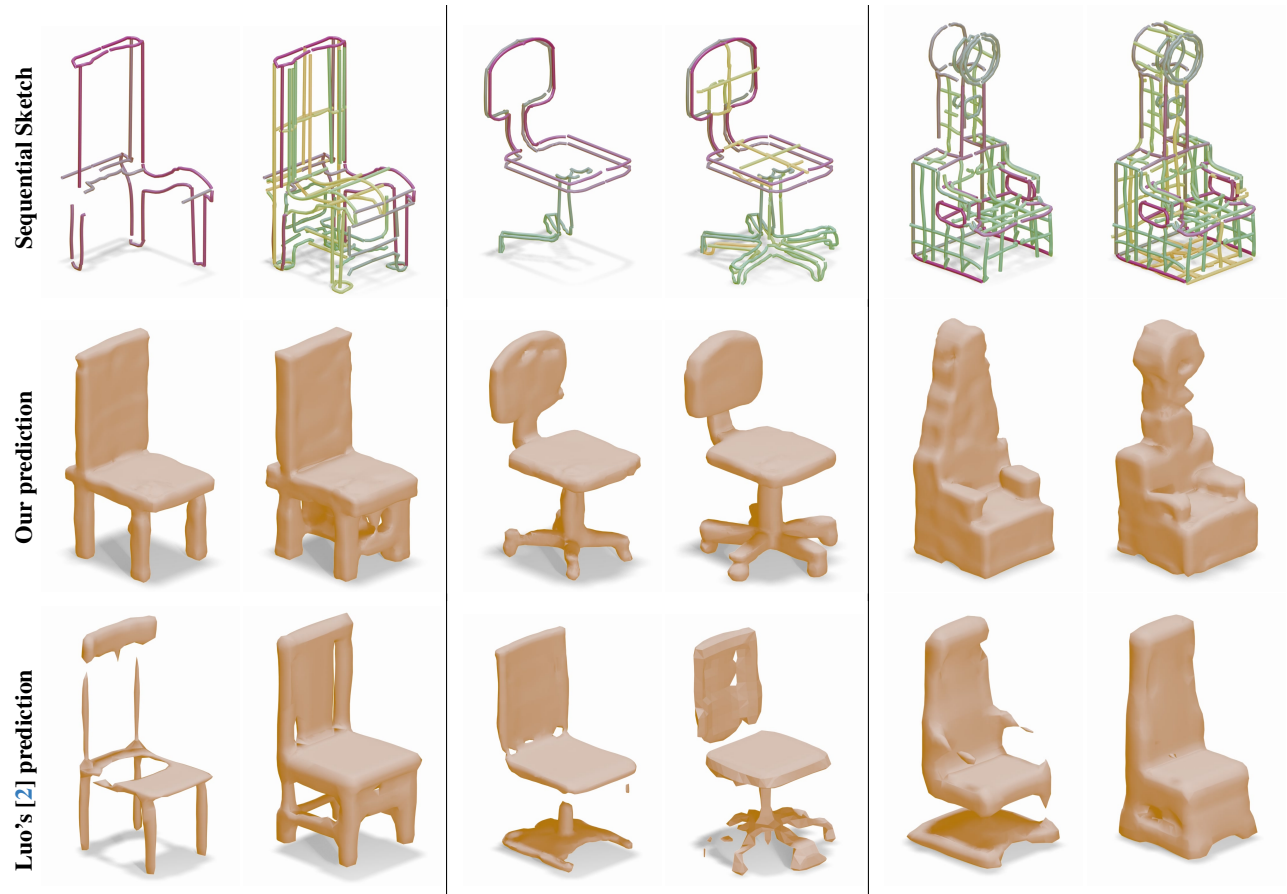
Figure A-1. **Sketch Completion Performance.** Performance remains high even when only partial sketches are provided.

We report completion results in Fig. A-2. Even when given only 25–50% of the original stroke sequence, the model infers coherent geometry and progressively refines the structure as more strokes are revealed, confirming its strong internal shape priors and robustness to missing sketch information. In particular, the model is able to exploit the geometry of shapes to complete un-sketched parts, such as missing chair legs.

As reported in Fig. A-1, our model is able to reconstruct faithful shapes even from partial inputs, clearly outperforming point cloud–based baselines. Interestingly, our model is able to reach near-maximum performance with only the first half of the drawn points. This trend reflects how annotators typically first draw global outlines of the shape before adding finer details.

A-3. Impact of Sketch Snapping

Our surface-snapping tool helps users draw geometrically accurate sketches directly on reference shapes. Because snapped sketches adhere to the underlying surface and remove user-induced alignment noise, they provide cleaner



(a) 25% sketch. (b) 100% sketch. (c) 50% sketch. (d) 100% sketch. (e) 75% sketch. (f) 100% sketch.

Figure A-2. **Sketch Completion Results.** Our model infers coherent 3D shapes even from highly partial sketches. As more strokes are provided, reconstructions become increasingly detailed and faithful to the target geometry.

supervision during training and yield more reliable evaluations. A natural concern, however, is whether models trained on such clean, snapped sketches might overfit to this idealized scenario and fail to generalize to sketches produced in the wild, without snapping assistance.

To assess this, we evaluate our model on sketches drawn without snapping and present qualitative results in Fig. A-3. Unsnapped sketches are visibly less precise and often contain wobbling or local distortions. Despite this domain shift, our model still produces convincing and geometrically faithful shapes, demonstrating strong robustness to deviations from perfectly aligned input sketches.

A-4. Evaluation on Free-Hand Sketches

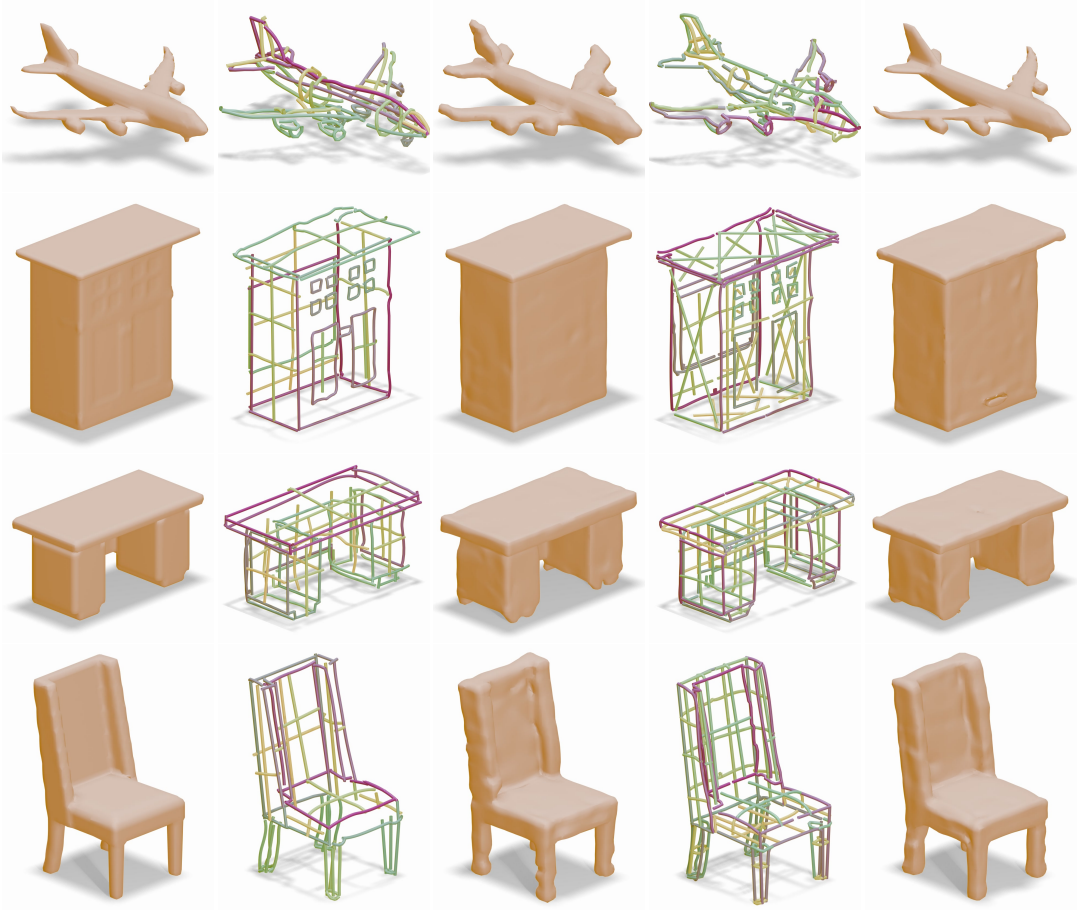
We provide additional illustrations for reconstruction of sketches drawn without references in Fig. A-4. We observe that despite clear stylistic and geometric differences from the training sketches, our model produces coherent, detailed, and semantically meaningful shapes that align well with the intent expressed in the free-hand inputs. By contrast, Luo *et al*. [2] generalizes poorly and tends to overfit to certain shape priors. These findings indicate that the model has learned a robust sketch-to-shape mapping rather than overfitting to the constraints of reference-guided drawing.

A-5. Evaluation on Unseen Classes

A related concern is whether a model trained exclusively on ShapeNet [1] categories truly learns a sketch-to-geometry mapping, or whether it merely exploits memorized class-specific priors. If the latter were the case, it should struggle when faced with sketches depicting objects outside the training categories.

To investigate this, we evaluate our model on sketches of *unseen categories* and *unseen shape collections*. Annotators produced VR sketches from ShapeNet classes excluded from training, as well as from the ModelNet dataset [3]. Representative results are shown in Fig. A-5.

Overall, the model generalizes surprisingly well: for many unseen categories, the generated shapes are coherent, structurally consistent, and aligned with the intent of



(a) GT shape. (b) Unsnapped sketch. (c) Our prediction. (d) Snapped sketch. (e) Our prediction.

Figure A-3. **Shape Generation from Sketches Without Snapping.** Sketches drawn without our snapping tool are noticeably noisier and less geometrically accurate, but our model can still generate coherent and plausible 3D shapes from them.

the sketch—despite never encountering such objects during training. However, the influence of learned priors remains visible in edge cases; for example, a sketched truck or bed may be reconstructed as an empty table-like structure, or a toilet may be reconstructed with a closed lid as a chair-like structure, reflecting the dominance of furniture categories in the training set.

These results indicate that the model has indeed learned a meaningful sketch-to-shape mapping that transfers across datasets and categories, while also revealing the limits of its current shape diversity and the role of priors when sketch evidence is sparse or ambiguous.

A-6. Impact of Sketcher Expertise

To evaluate the impact of the expertise of the sketcher, we evaluated sketches of 100 shapes drawn by both expert annotators and beginners. Non-expert sketches yield higher reconstruction accuracy (72.2 ± 2.6 vs. 66.3 ± 6.4 F-score),. We hypothesis that this is because experts introduce greater abstraction while beginners trace geometry more faithfully.

A-7. Precision/Performance Tradeoff

In the main paper, we reported results using 100 DDIM steps, which yield the best reconstruction quality but account for over 99% of the inference time. This setting results in a latency of roughly 6 seconds per sketch, which may be impractical in interactive design scenarios.

To assess whether fewer sampling steps provide a better speed–accuracy compromise, we evaluate our model with 10, 25, 50, and 100 DDIM steps. As shown in Tab. A-2, performance remains remarkably stable even with as few as 10 steps, while inference becomes over $3\times$ faster. This indicates that interactive or real-time use cases can run with drastically reduced sampling budgets at minimal loss in quality.

A-8. Additional Qualitative Results

We present additional comparisons with Luo *et al*. [2] and LAS-Diffusion [4] in Fig. A-6. Although LAS-Diffusion

Figure A-4. **Shape Generation from Free-Hand Sketches.** Compared with Luo *et al*. [2], Our model generalizes well to free-hand sketches drawn without any reference shape for airplanes, chairs/sofas, tables, and cabinets, producing detailed and plausible reconstructions that reflect the user's intent.

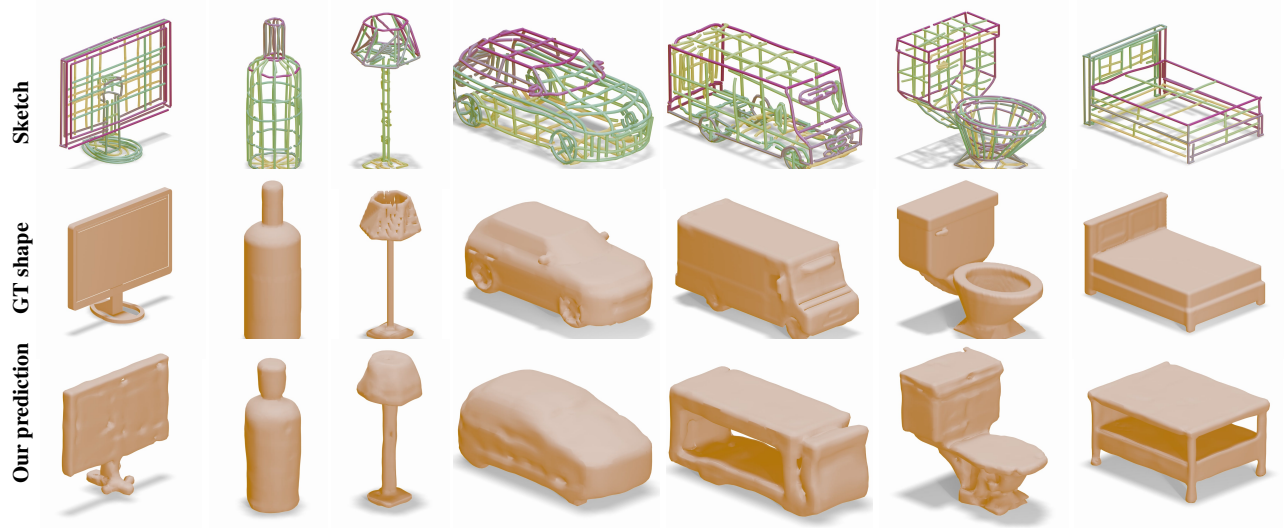


Figure A-5. **Shape Generation from Unseen Classes.** Results on sketches depicting object categories not present in the training data, including bottles, lamps, and cars from ShapeNet [1], and monitors, toilets, and beds from ModelNet [3]. Despite the domain shift, our model generally produces plausible shapes aligned with the sketch intent. However, some predictions reveal an overreliance on learned priors: trucks or beds may be completed into table-like structures.

DDIM step	F-score ↑	CD $\times 1000$ ↓	time (samples/s)
10	69.24	5.04	2.26
25	69.70	4.82	3.06
50	69.74	4.89	4.47
100	69.80	4.78	6.33

Table A-2. **Performance/Speed Tradeoff.** Reconstruction accuracy remains stable even with a small number of DDIM steps, while inference becomes substantially faster (computed with a batch size of 1).

yields smooth surfaces, it struggles with occlusions and fails to generate complete geometry, consistent with its higher Chamfer errors. Our method produces more detailed, structurally accurate shapes across diverse sketches.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv:1512.03012*, 2015. 2, 5

[2] Ling Luo, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song, and Yulia Gryaditskaya. 3D VR sketch guided 3D shape prototyping and exploration. In *ICCV*, 2023. 1, 2, 3, 4, 6

[3] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2, 5

[4] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional SDF diffusion for controllable 3D shape generation. *ACM Transactions on Graphics (ToG)*, 2023. 3, 6

Figure A-6. **Additional Qualitative Illustrations.** Comparison between our method, Luo *et al*. [2], and LAS-Diffusion [4] on the real test set of VRSKETCH2SHAPE.