Hyperstroke: A Novel High-quality Stroke Representation for Assistive Artistic Drawing (Supplementary Material)

Haoyun Qin
University of Pennsylvania
Philadelphia, PA, USA
qhy@seas.upenn.edu

Jian Lin
Saint Francis University
Hong Kong, China
jlin@sfu.edu.hk

Hanyuan Liu
City University of Hong Kong
Hong Kong, China
hy.liu@cityu.edu.hk

Xueting Liu
Saint Francis University
Hong Kong, China
tliu@sfu.edu.hk

Chengze Li
Saint Francis University
Hong Kong, China
czli@sfu.edu.hk

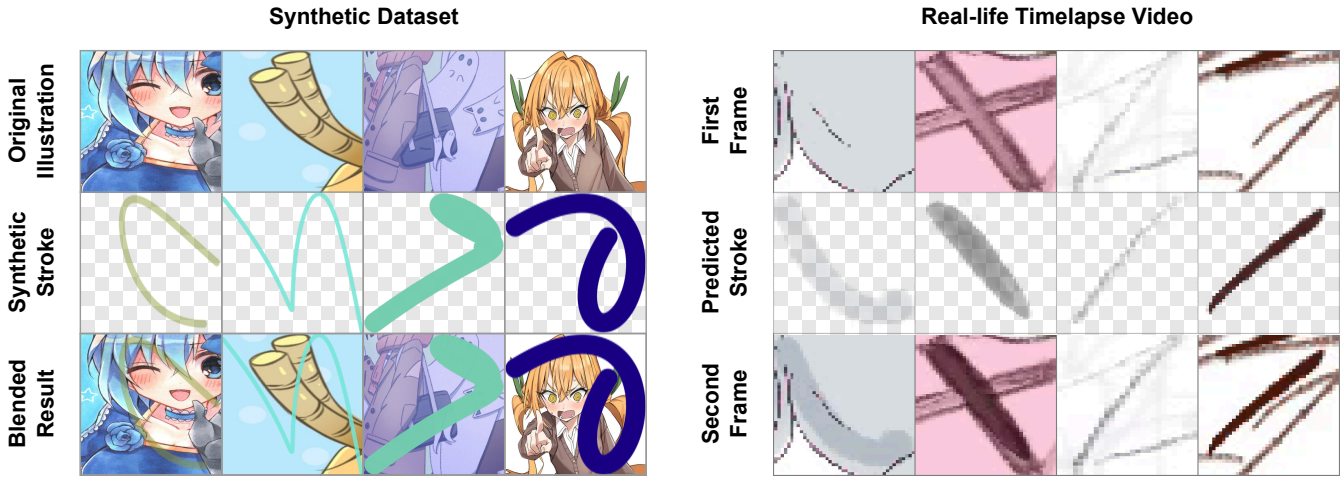

Figure 1: Data examples to train the Hyperstroke representation. The first group shows the data from synthetic dataset. From top to bottom are original illustrations, synthetic stroke images, and blended results. The supervision is conducted directly by the ground truth synthetic stroke. The second group demonstrates the data from real-life timelapse video, showing the previous frames in the frame pairs, the *predicted* stroke by our model (not part of the dataset), and the latter frames in the frame pairs, from the top to bottom accordingly. Here, the supervision is implicitly applied by the two frames.

1 Hyperstroke Dataset Showcase

As stated, the training data for hyperstroke consists of a synthetic dataset and data from real-life timelapse video. Figure 1 shows some samples from our dataset. Specifically, for the synthetic dataset, we left 30% of the strokes opaque while the other strokes have a uniform opacity sampled from 0.1 to 1.0, and the background images are obtained from [1].

2 Training Details

2.1 Hyperstroke Representation

Our VQ model employs a codebook of 8192 vocabularies, each with 256 dimensions of embedding. We trained the model using the mixed dataset consisting of 85,425 synthetic data samples and 74,286 real data samples in the form of frame pairs on a resolution of 256×256. The base learning rate is 4.5×10^{-6}, and is warmed up for 200 steps linearly at the beginning. The model is trained for 20 epochs on 4 NVIDIA A6000 GPUs for 35 hours at a total batch size of 32.

2.2 Assistive Sketch Generation

We evaluated the assistive sketch generation task on the *Quick, Draw!* dataset [2], which consists of temporal vector sketching of 345 categories. We first filter the dataset for sketches with stroke number ranging from 3 to 15, subsample the dataset by 1/5, and then render each sketch in black color with random stroke width on arbitrary canvas positions, resulting the final dataset with 43,776,398 strokes from 7,049,475 sketches. To employ the transformer model on the new dataset, a new VQ model is trained. Under this setting, we make the following changes: we adopt the codebook size of 2048, chose 2×10^{-7} as the base learning rate, and trained the model on a resolution of 128×128 for 2 epochs on 8 NVIDIA A100 GPUs for 35 hours at a total batch size of 1024. The downsampling factor of the VQ model is 16×, and therefore a hyperstroke consists of $4 + (128/16)^2 = 68$ tokens. For the transformer model, we adopt

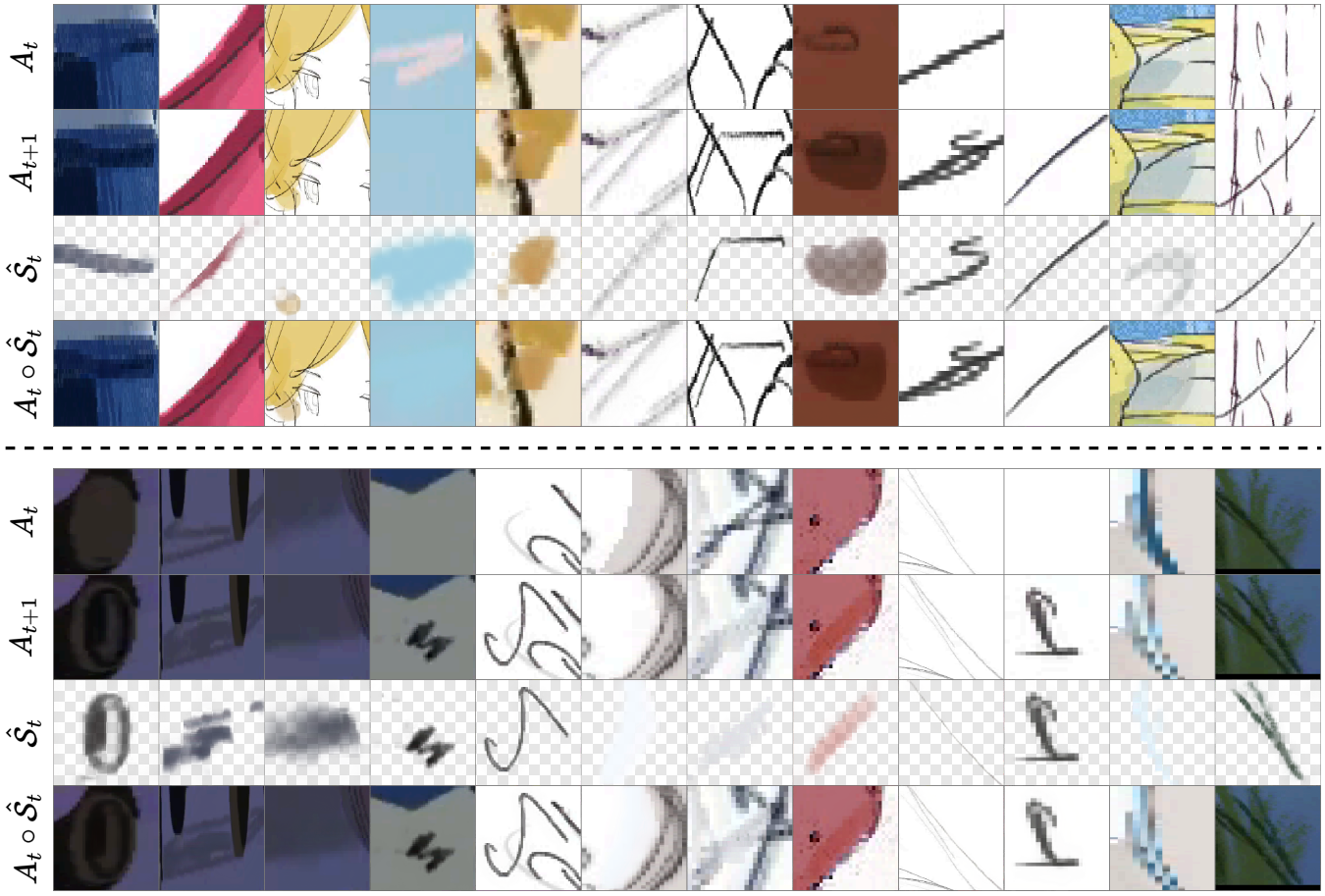

Figure 2: Hyperstroke model result on real-life timelapse drawing data. For each group, the four rows from the top to the bottom stand for the previous frame, the latter frame, the predicted strokes between the two frames, and finally the blended result of the predicted strokes onto the initial frames.

GPT-2 (345M) [3] as the decoder, a pretrained Vision Transformer (ViT)[1] as the canvas encoder, and a pretrained CLIP model[2] as the control encoder. Here, we condition the generation based on the category text of each sketch, and the context length is 12 strokes, i.e. $1 + 12 \times 68 = 817$ tokens. The learning rate is 5×10^{-4} and we employed learning warmup and annealing. We freeze the weights of encoders during the training, and the model is trained for 1 epoch on 8 NVIDIA A100 GPUs for 3 days at a total batch size of 1024.

3 Additional Experiment Results

3.1 Hyperstroke Reconstruction

Figure 3 and Figure 2 show results on synthetic dataset and real-life drawing timelapse data accordingly.

3.2 Assistive Sketch Generation

Figure 4 shows generation results conditioned on blank canvas and seen text categories. Figure 5 demonstrates the results where the canvas is half-way finished. We also tested our model on *unseen* text conditions beyond the 345 text categories the model is trained on as shown in Figure 6. The model demonstrates the capability of extrapolation to some extent, where it can guess the overall shape and feel of unseen data in some cases.

[1] https://huggingface.co/google/vit-base-patch16-224-in21k
[2] https://huggingface.co/openai/clip-vit-base-patch32

References

[1] Nyanko Devs. [n. d.]. Danbooru2023 Dataset. https://huggingface.co/datasets/nyanko7/danbooru2023. Accessed: 2024-10-02.
[2] David Ha and Douglas Eck. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477* (2017).
[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

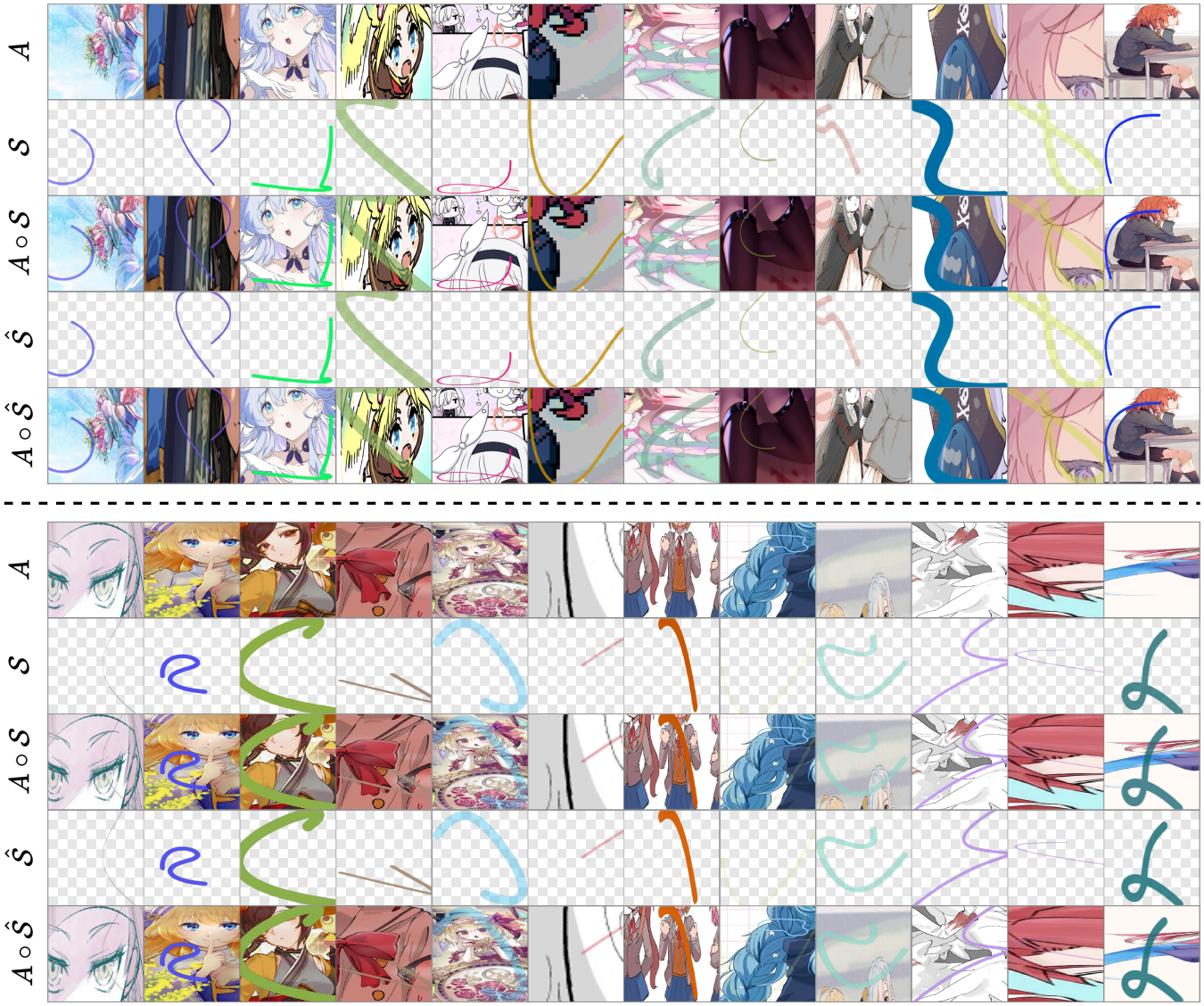

Figure 3: Hyperstroke model result on synthetic dataset. For each group, the five rows from the top to the bottom stand for the original cropped illustration, the generated ground truth stroke images, the blended illustration by the ground truth, the predicted strokes between the two frames, and finally the blended result of the predicted strokes.

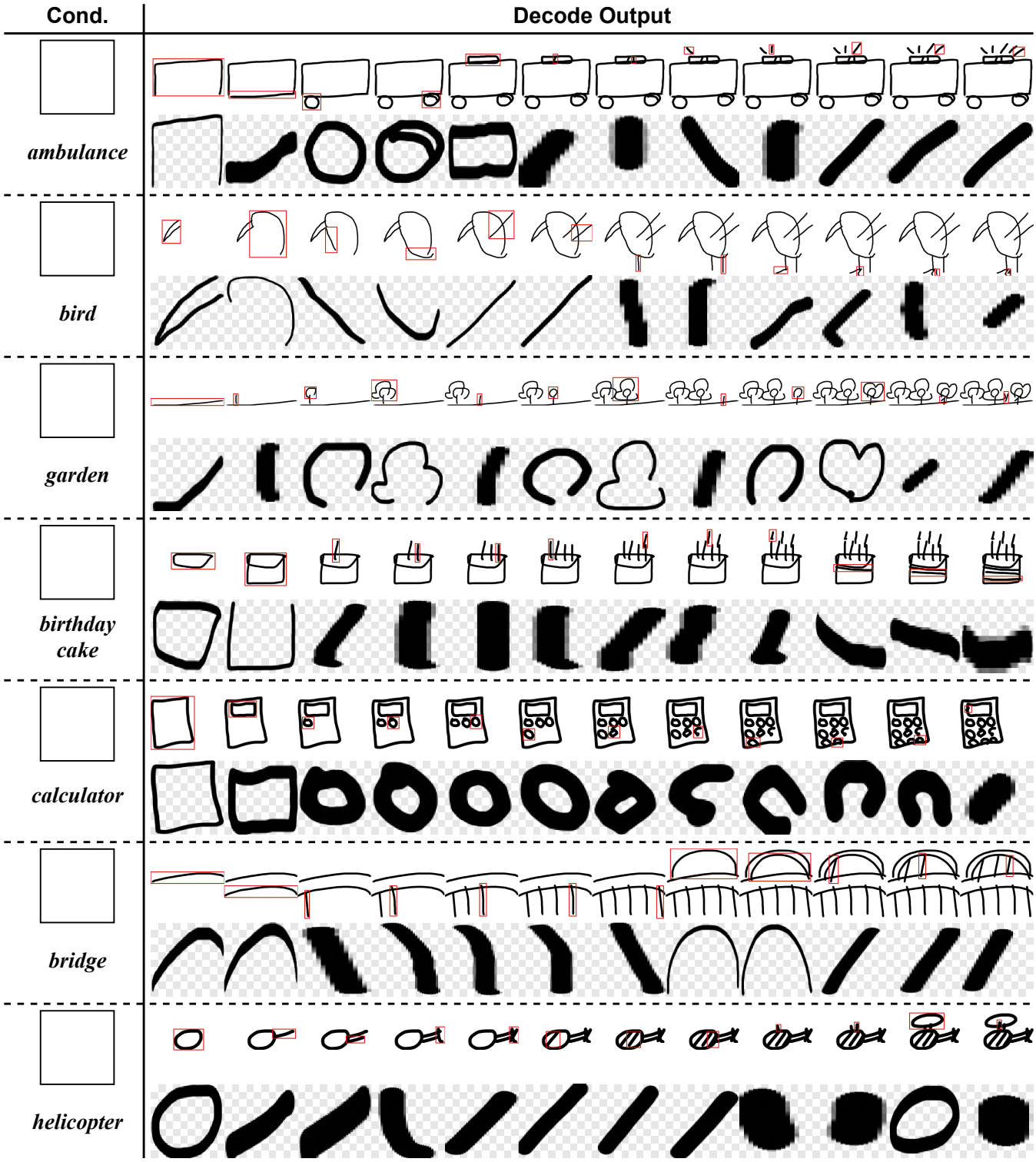

Figure 4: Results on assistive sketch generation from blank canvas, conditioned on seen text categories.

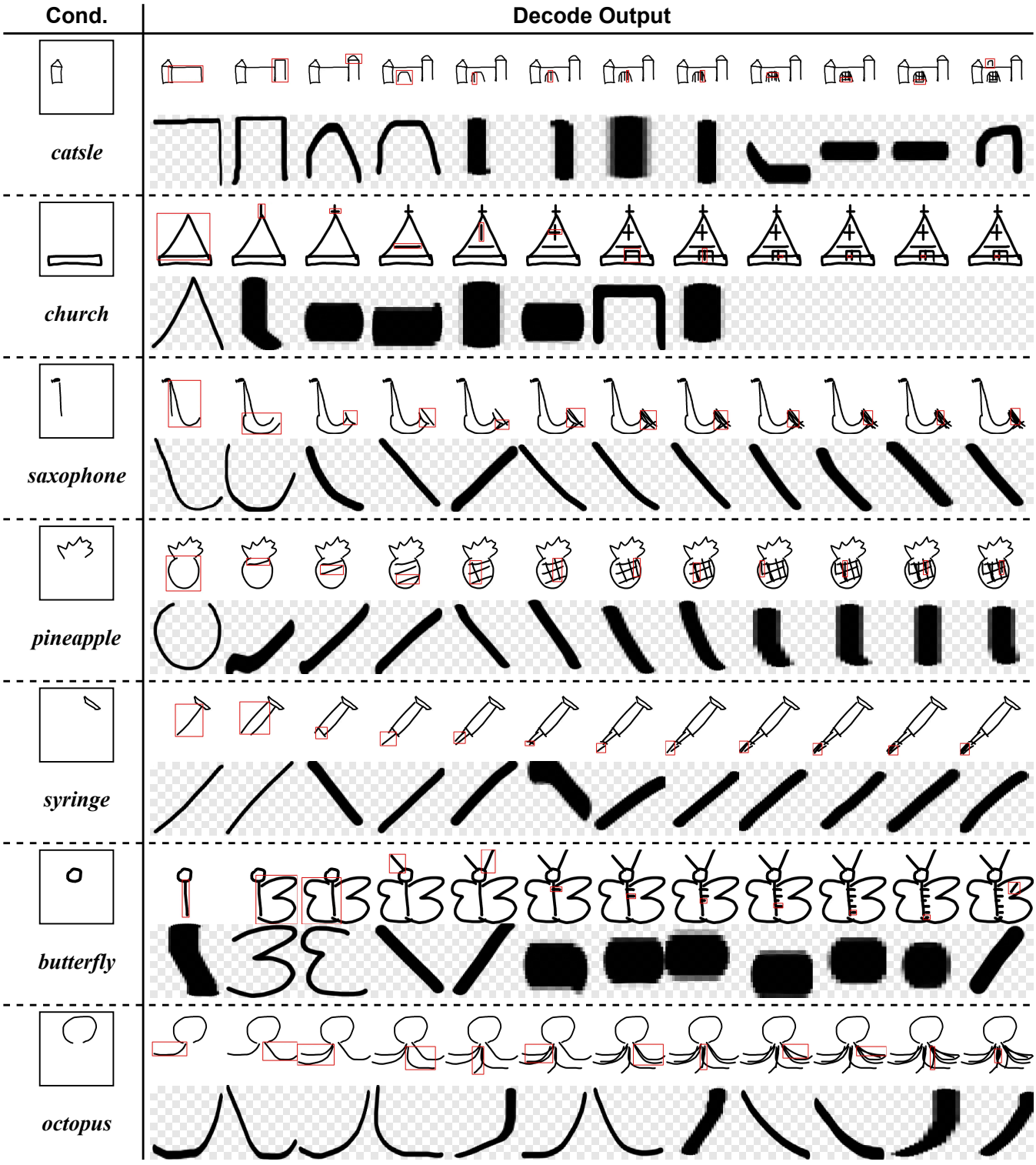

Figure 5: Results on assistive sketch generation, conditioned the raster canvas images and seen text categories.

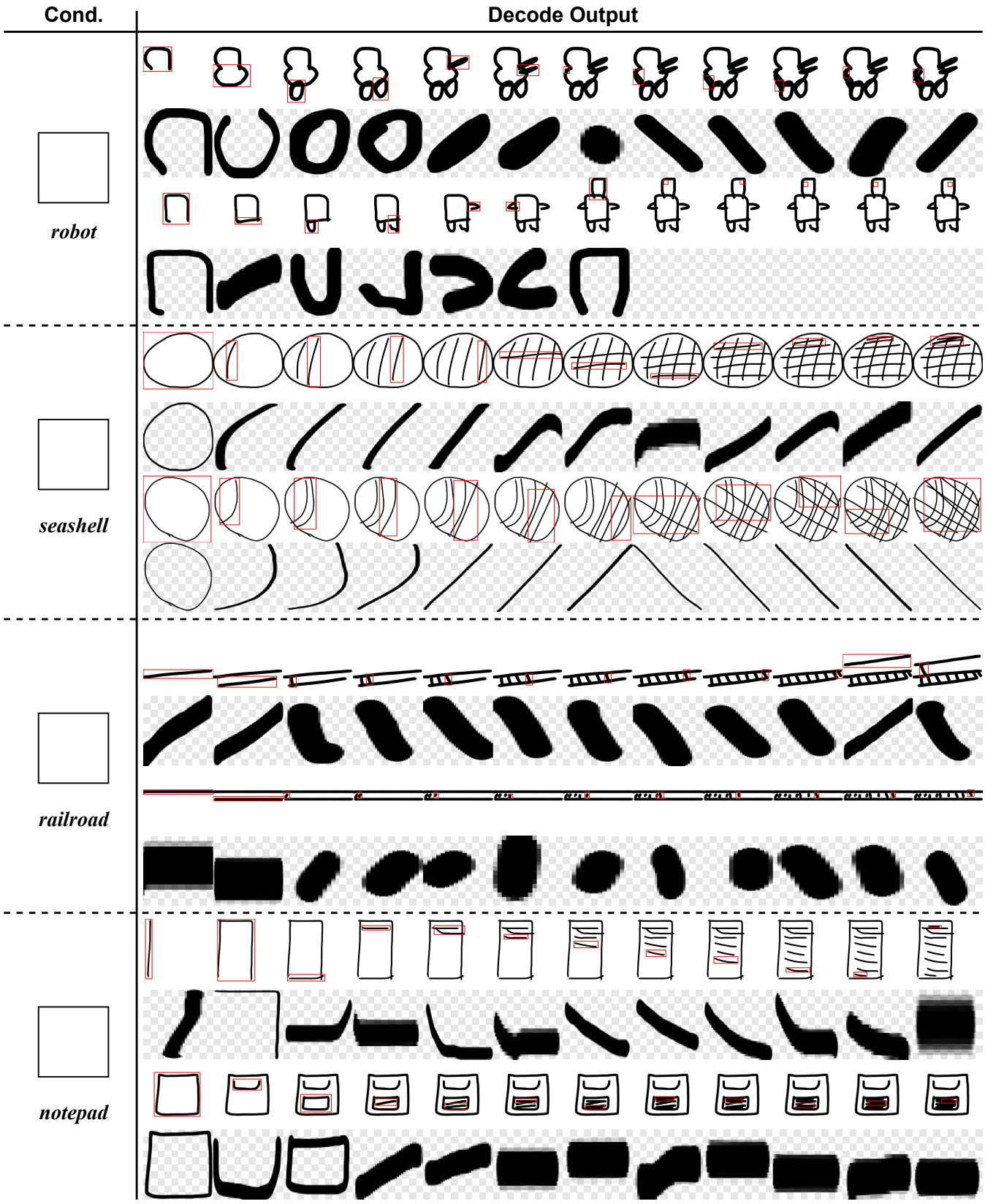

Figure 6: Results on assistive sketch generation from blank canvas, conditioned on unseen text categories.