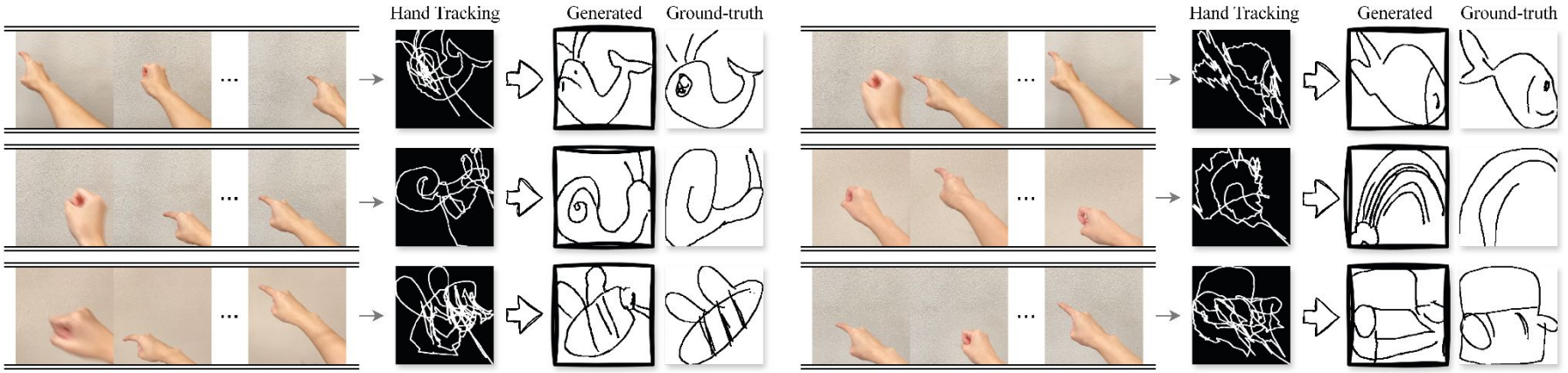
AirSketch: Generative Motion to Sketch

Hui Xian Grace Lim, Xuanming Cui, Yogesh S Rawat, Ser-Nam Lim

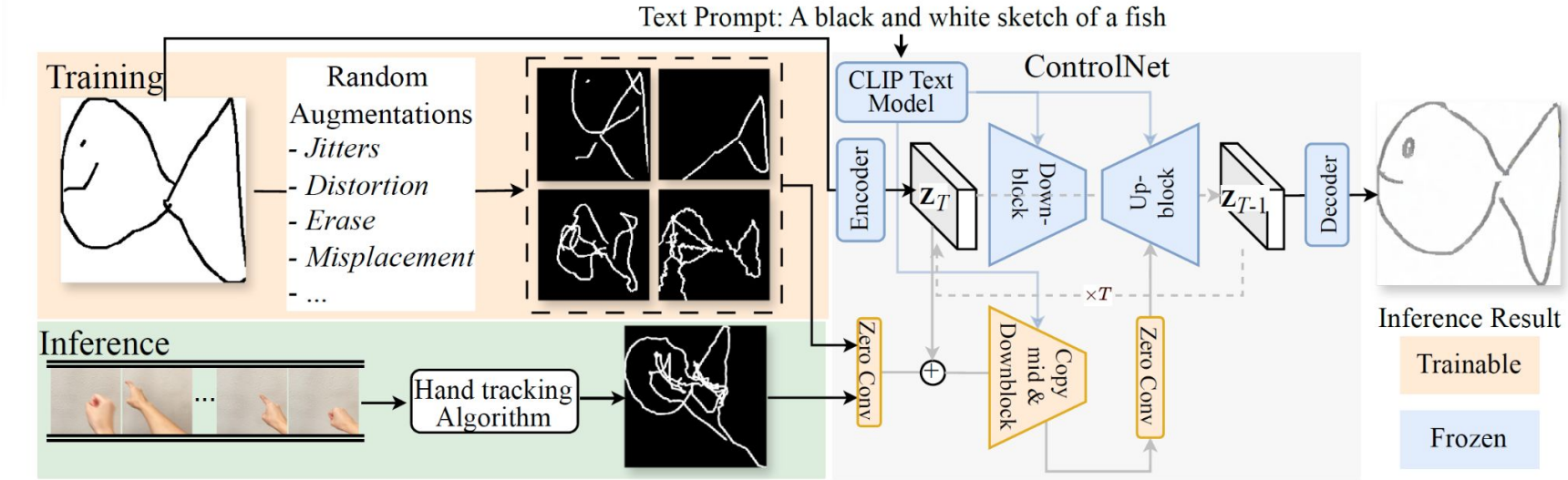
Freeform, gesture based illustration



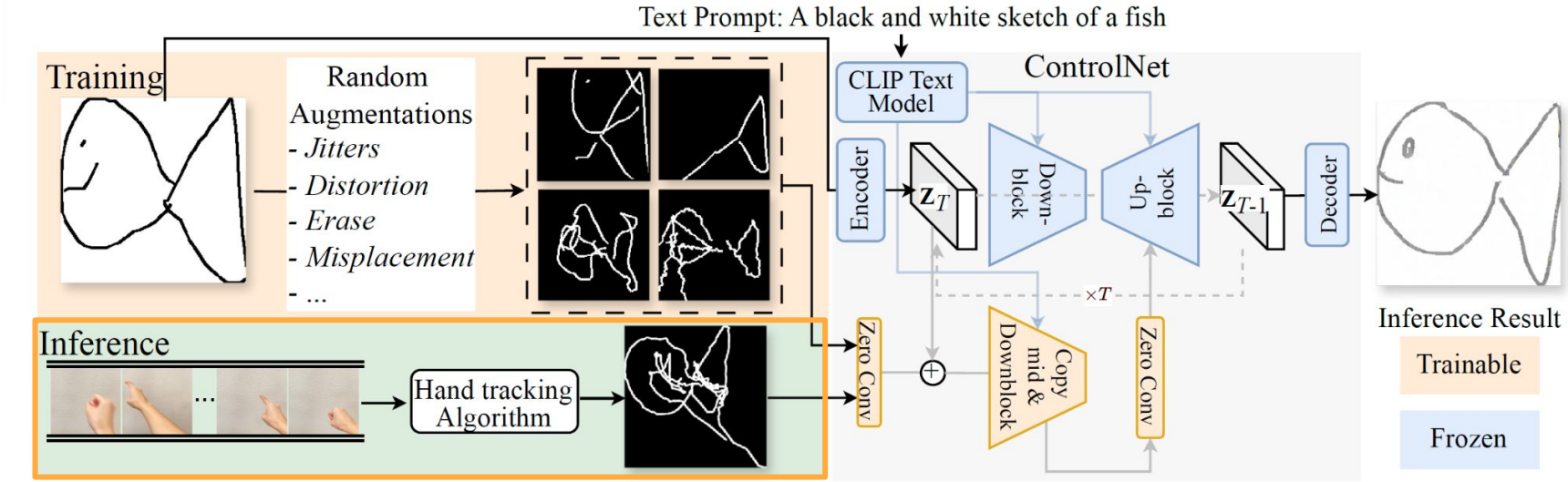
Goal: Markerless, hardware-agnostic framework for hand motion-based illustration.

Task: Generate visually coherent and aesthetic sketches directly from hand motions captured with a simple RGB camera.

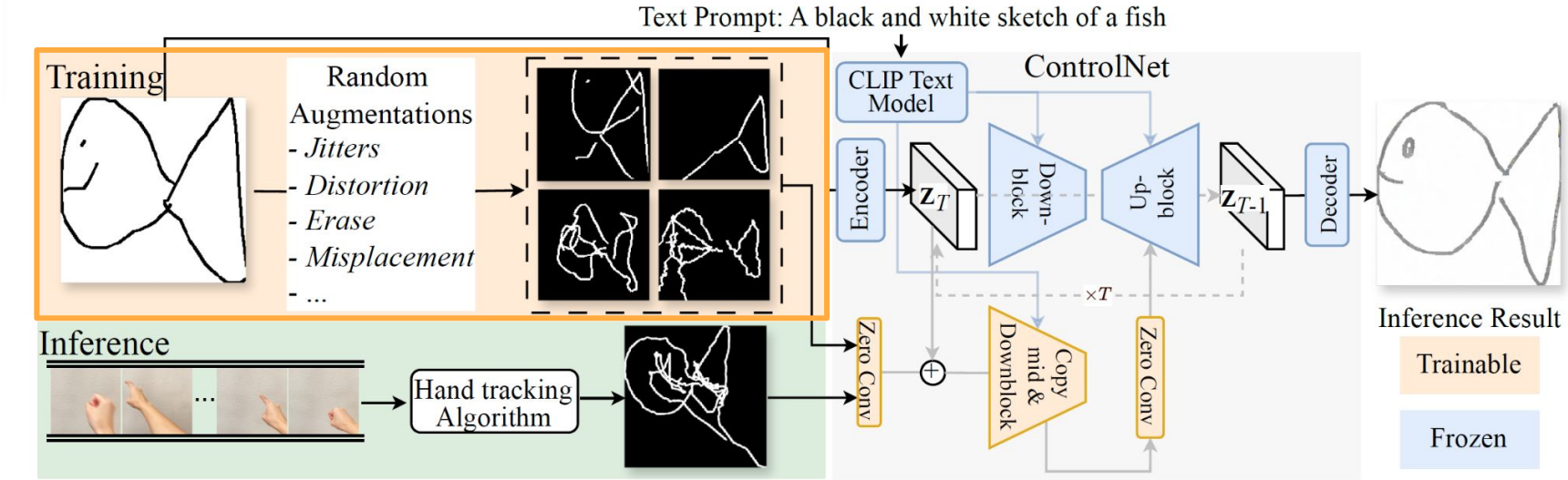
Approach


Text Prompt: A black and white sketch of a fish
Training
Random Augmentations
- Jitters
- Distortion
- Erase
- Misplacement
- ...
CLIP Text Model
ControlNet
Encoder
z_T
Down-block
Up-block
z_{T-1}
Decoder
Inference Result
$\times T$
Inference
Hand tracking Algorithm
Zero Conv
Copy mid & Downblock
Zero Conv
Trainable
Frozen

Approach


Text Prompt: A black and white sketch of a fish
Training
Random Augmentations
- Jitters
- Distortion
- Erase
- Misplacement
- ...
Inference
Hand tracking Algorithm
CLIP Text Model
ControlNet
Encoder
z_T
Down-block
Up-block
z_{T-1}
Decoder
$\times T$
Zero Conv
Copy mid & Downblock
Zero Conv
Inference Result
Trainable
Frozen

Approach

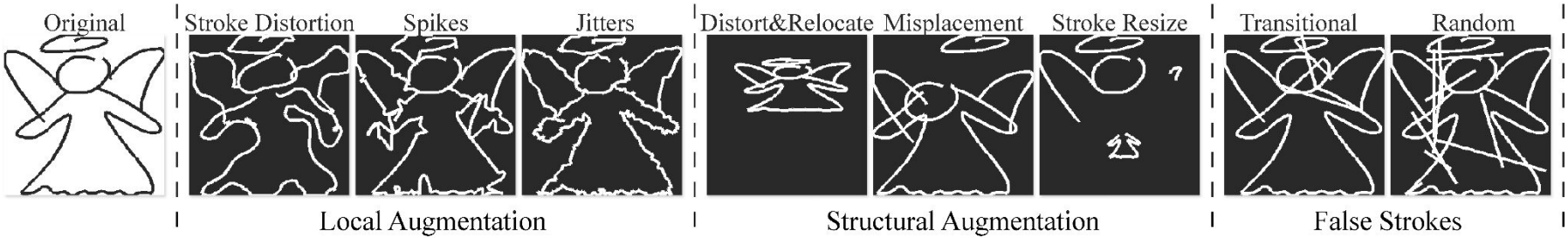

Text Prompt: A black and white sketch of a fish
Training
Random Augmentations
- Jitters
- Distortion
- Erase
- Misplacement
- ...
Inference
Hand tracking Algorithm
CLIP Text Model
ControlNet
Encoder
z_T
Down-block
Up-block
z_{T-1}
Decoder
$\times T$
Zero Conv
Copy mid & Downblock
Zero Conv
Inference Result
Trainable
Frozen

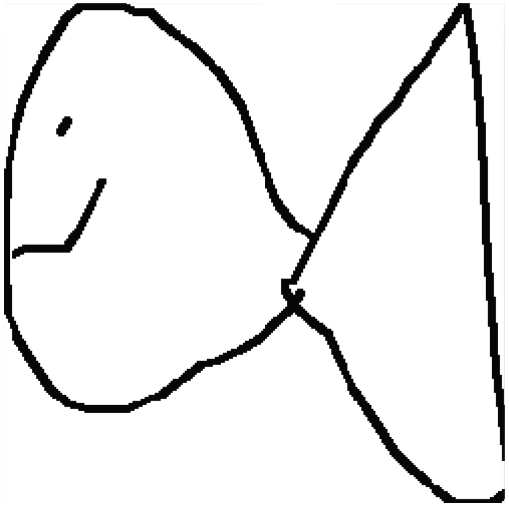
Representing input artifacts with augmentations

Local artifacts: jitters, stroke-wise distortion, random spikes

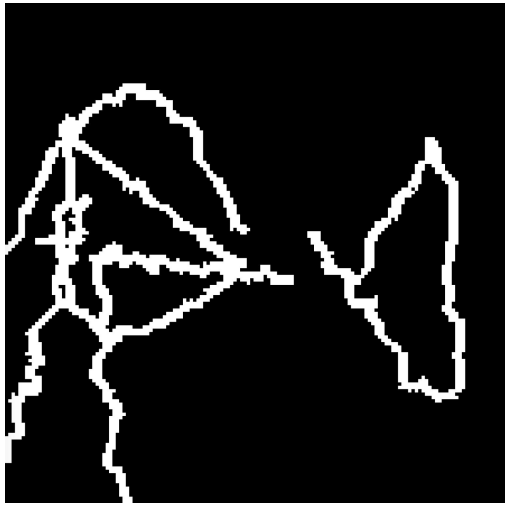
Structural artifacts: sketch-level distortion, incorrect stroke size, misplacement

False strokes: entering/exiting canvas, transition between strokes, hesitation





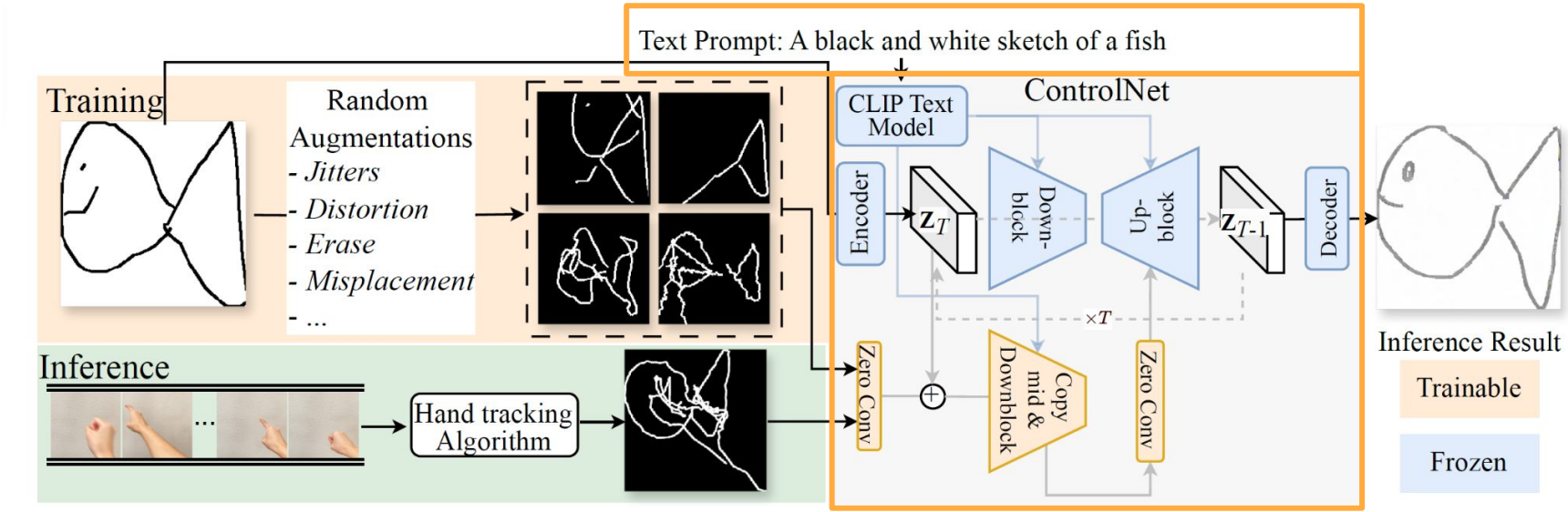
Ground truth sketch



Augmented ground truth sketch

Example result from hand tracking

Approach

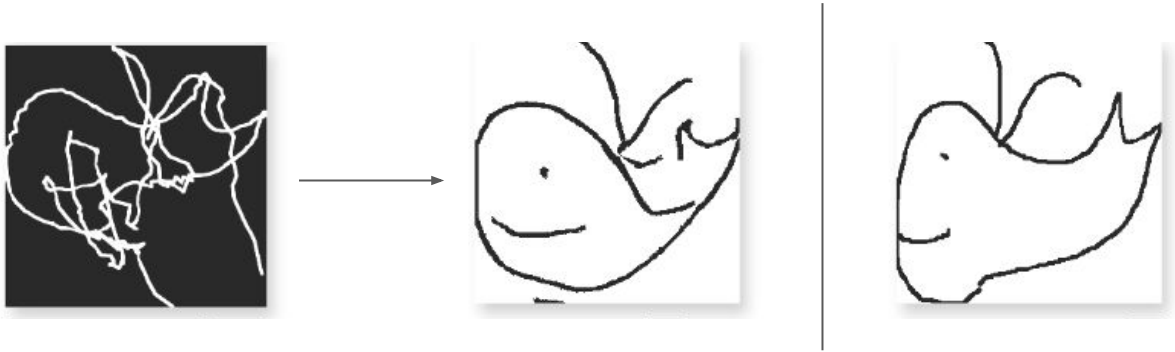

Text Prompt: A black and white sketch of a fish
Training
Random Augmentations
- Jitters
- Distortion
- Erase
- Misplacement
- ...
Inference
Hand tracking Algorithm
ControlNet
CLIP Text Model
Encoder
z_T
Down-block
Up-block
z_{T-1}
Decoder
$\times T$
Zero Conv
Copy mid & Downblock
Zero Conv
Inference Result
Trainable
Frozen

Recovering original sketch from chaotic input

Input: Noisy input image, prompt: “A black and white sketch of a <noun>”.

- ControlNet: input image is structurally identical to the desired output

Output: Original, undistorted sketch.

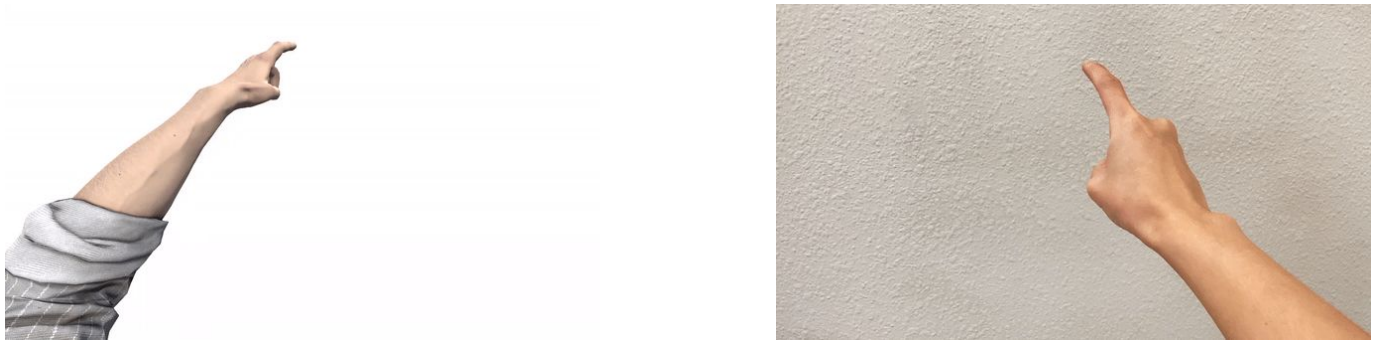


Hand tracking result and generated image, reference image

Hand Motion Data Collection

Synthetic: Animated in Unity, guided by Quick, Draw! dataset: 50 categories, 100 videos each

Real: Human user attempts to replicate samples from Quick, Draw! dataset: 50 categories, 10 videos each



Clips from synthetic (left) and real (right) hand motion datasets

Table 1: Results on the similarity between generated and ground-truth sketches from Quick, Draw! dataset. "Tracking" refers to hand tracking images, and "Gen." refers to generated images. "w/ Aug." refers to whether sketch augmentations have been applied. "CLIP I2I/I2T" refers to CLIP Image-to-Image/Image-to-Text similarity.

	Dataset	Backbone	w/ Aug.	SSIM (↑)	CD (↓)	LPIPS (↓)	CLIP I2I (↑)	CLIP I2T (↑)
Seen Categories								
Tracking	synth.	–	–	0.59	20.12	0.36	0.80	0.22
Gen.	synth.	SDXL	✗	0.59	20.11	0.37	0.79	0.23
Gen.	synth.	SD1.5	✓	0.60	17.98	0.35	0.80	0.26
Gen.	synth.	SDXL	✓	**0.64**	**17.39**	**0.33**	**0.85**	**0.28**
Tracking	real	–	–	0.55	32.36	0.42	0.76	0.21
Gen.	real	SDXL	✗	0.55	31.99	0.41	0.79	0.21
Gen.	real	SD1.5	✓	0.59	27.59	0.38	0.80	0.27
Gen.	real	SDXL	✓	**0.64**	**25.46**	**0.36**	**0.84**	**0.29**
Unseen Categories								
Tracking	synth.	–	–	0.59	20.47	0.36	0.80	0.22
Gen.	synth.	SDXL	✗	0.59	20.32	0.35	0.81	0.22
Gen.	synth.	SD1.5	✓	0.60	17.50	0.35	0.80	0.26
Gen.	synth.	SDXL	✓	**0.64**	**17.27**	**0.34**	**0.85**	**0.27**
Tracking	real	–	–	0.54	33.92	0.42	0.76	0.21
Gen.	real	SDXL	✗	0.55	33.53	0.41	0.78	0.21
Gen.	real	SD1.5	✓	0.61	27.67	0.38	0.80	0.27
Gen.	real	SDXL	✓	**0.63**	**24.26**	**0.38**	**0.85**	**0.28**

Table 1: Results on the similarity between generated and ground-truth sketches from Quick, Draw! dataset. "Tracking" refers to hand tracking images, and "Gen." refers to generated images. "w/ Aug." refers to whether sketch augmentations have been applied. "CLIP I2I/I2T" refers to CLIP Image-to-Image/Image-to-Text similarity.

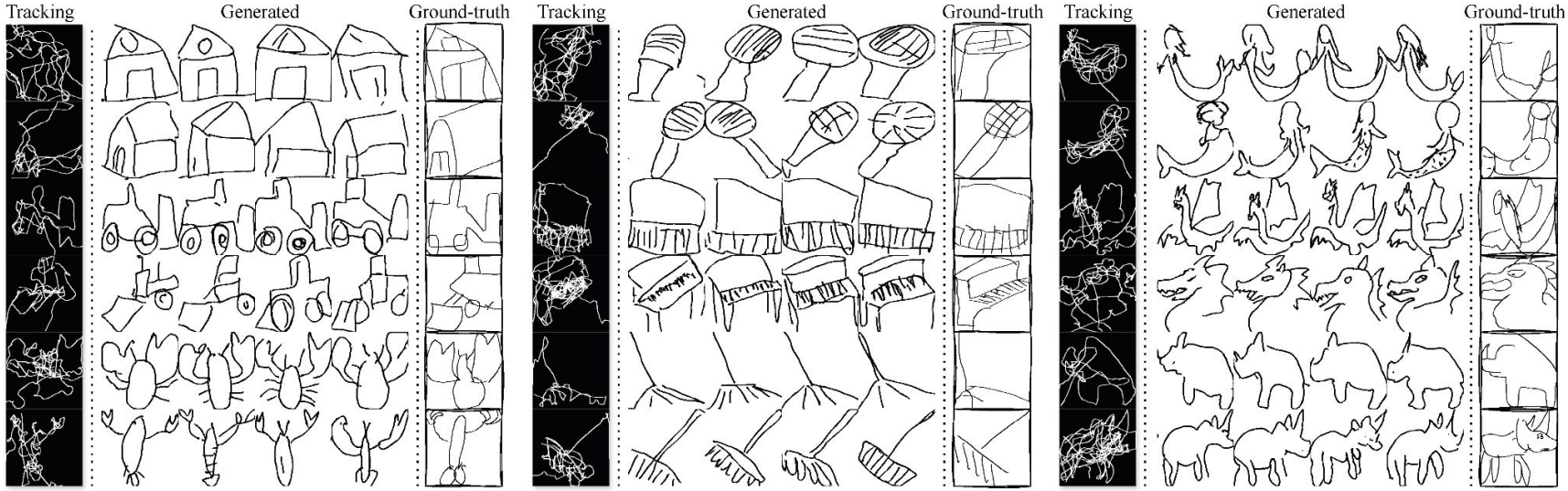
	Dataset	Backbone	w/ Aug.	SSIM (↑)	CD (↓)	LPIPS (↓)	CLIP I2I (↑)	CLIP I2T (↑)
Seen Categories								
Tracking	synth.	–	–	0.59	20.12	0.36	0.80	0.22
Gen.	synth.	SDXL	✗	0.59	20.11	0.37	0.79	0.23
Gen.	synth.	SD1.5	✓	0.60	17.98	0.35	0.80	0.26
Gen.	synth.	SDXL	✓	**0.64**	**17.39**	**0.33**	**0.85**	**0.28**
Tracking	real	–	–	0.55	32.36	0.42	0.76	0.21
Gen.	real	SDXL	✗	0.55	31.99	0.41	0.79	0.21
Gen.	real	SD1.5	✓	0.59	27.59	0.38	0.80	0.27
Gen.	real	SDXL	✓	**0.64**	**25.46**	**0.36**	**0.84**	**0.29**
Unseen Categories								
Tracking	synth.	–	–	0.59	20.47	0.36	0.80	0.22
Gen.	synth.	SDXL	✗	0.59	20.32	0.35	0.81	0.22
Gen.	synth.	SD1.5	✓	0.60	17.50	0.35	0.80	0.26
Gen.	synth.	SDXL	✓	**0.64**	**17.27**	**0.34**	**0.85**	**0.27**
Tracking	real	–	–	0.54	33.92	0.42	0.76	0.21
Gen.	real	SDXL	✗	0.55	33.53	0.41	0.78	0.21
Gen.	real	SD1.5	✓	0.61	27.67	0.38	0.80	0.27
Gen.	real	SDXL	✓	**0.63**	**24.26**	**0.38**	**0.85**	**0.28**

Table 1: Results on the similarity between generated and ground-truth sketches from Quick, Draw! dataset. "Tracking" refers to hand tracking images, and "Gen." refers to generated images. "w/ Aug." refers to whether sketch augmentations have been applied. "CLIP I2I/I2T" refers to CLIP Image-to-Image/Image-to-Text similarity.

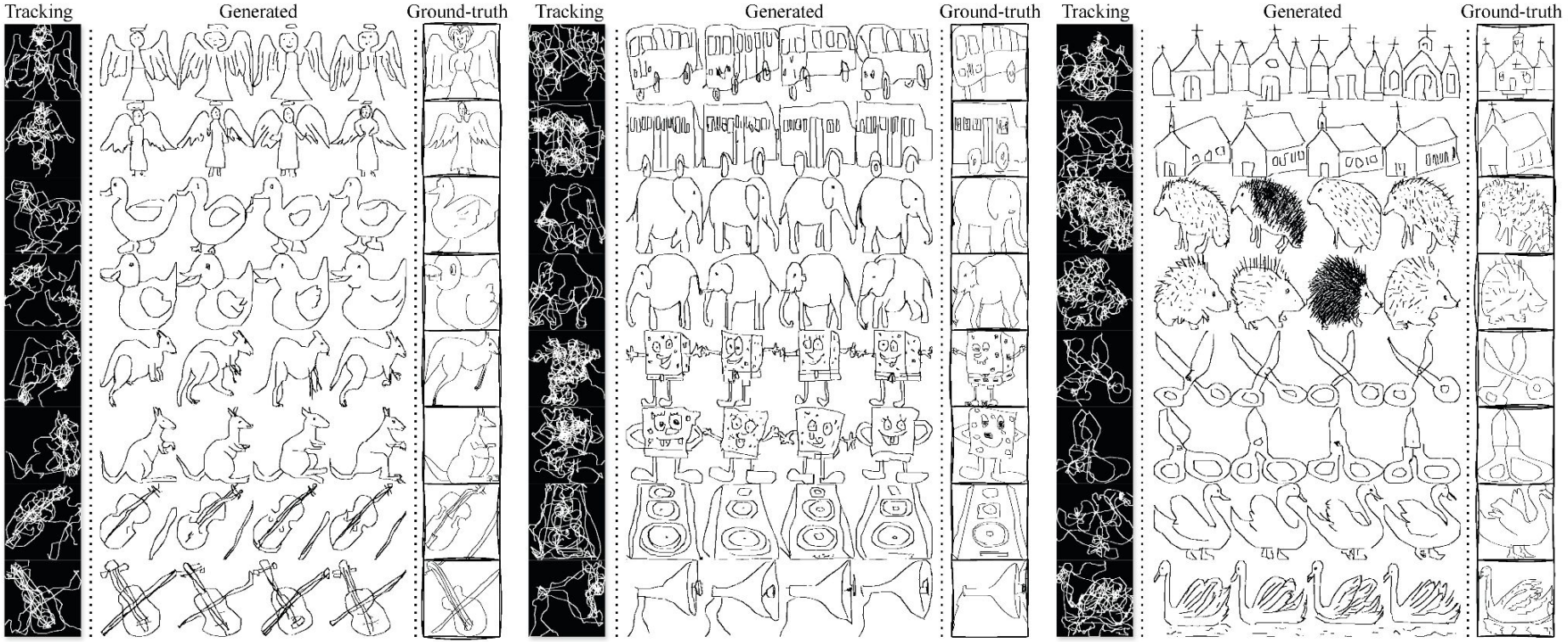
	Dataset	Backbone	w/ Aug.	SSIM (↑)	CD (↓)	LPIPS (↓)	CLIP I2I (↑)	CLIP I2T (↑)
Seen Categories								
Tracking	synth.	–	–	0.59	20.12	0.36	0.80	0.22
Gen.	synth.	SDXL	✗	0.59	20.11	0.37	0.79	0.23
Gen.	synth.	SD1.5	✓	0.60	17.98	0.35	0.80	0.26
Gen.	synth.	SDXL	✓	**0.64**	**17.39**	**0.33**	**0.85**	**0.28**
Tracking	real	–	–	0.55	32.36	0.42	0.76	0.21
Gen.	real	SDXL	✗	0.55	31.99	0.41	0.79	0.21
Gen.	real	SD1.5	✓	0.59	27.59	0.38	0.80	0.27
Gen.	real	SDXL	✓	**0.64**	**25.46**	**0.36**	**0.84**	**0.29**
Unseen Categories								
Tracking	synth.	–	–	0.59	20.47	0.36	0.80	0.22
Gen.	synth.	SDXL	✗	0.59	20.32	0.35	0.81	0.22
Gen.	synth.	SD1.5	✓	0.60	17.50	0.35	0.80	0.26
Gen.	synth.	SDXL	✓	**0.64**	**17.27**	**0.34**	**0.85**	**0.27**
Tracking	real	–	–	0.54	33.92	0.42	0.76	0.21
Gen.	real	SDXL	✗	0.55	33.53	0.41	0.78	0.21
Gen.	real	SD1.5	✓	0.61	27.67	0.38	0.80	0.27
Gen.	real	SDXL	✓	**0.63**	**24.26**	**0.38**	**0.85**	**0.28**

Table 1: Results on the similarity between generated and ground-truth sketches from Quick, Draw! dataset. "Tracking" refers to hand tracking images, and "Gen." refers to generated images. "w/ Aug." refers to whether sketch augmentations have been applied. "CLIP I2I/I2T" refers to CLIP Image-to-Image/Image-to-Text similarity.

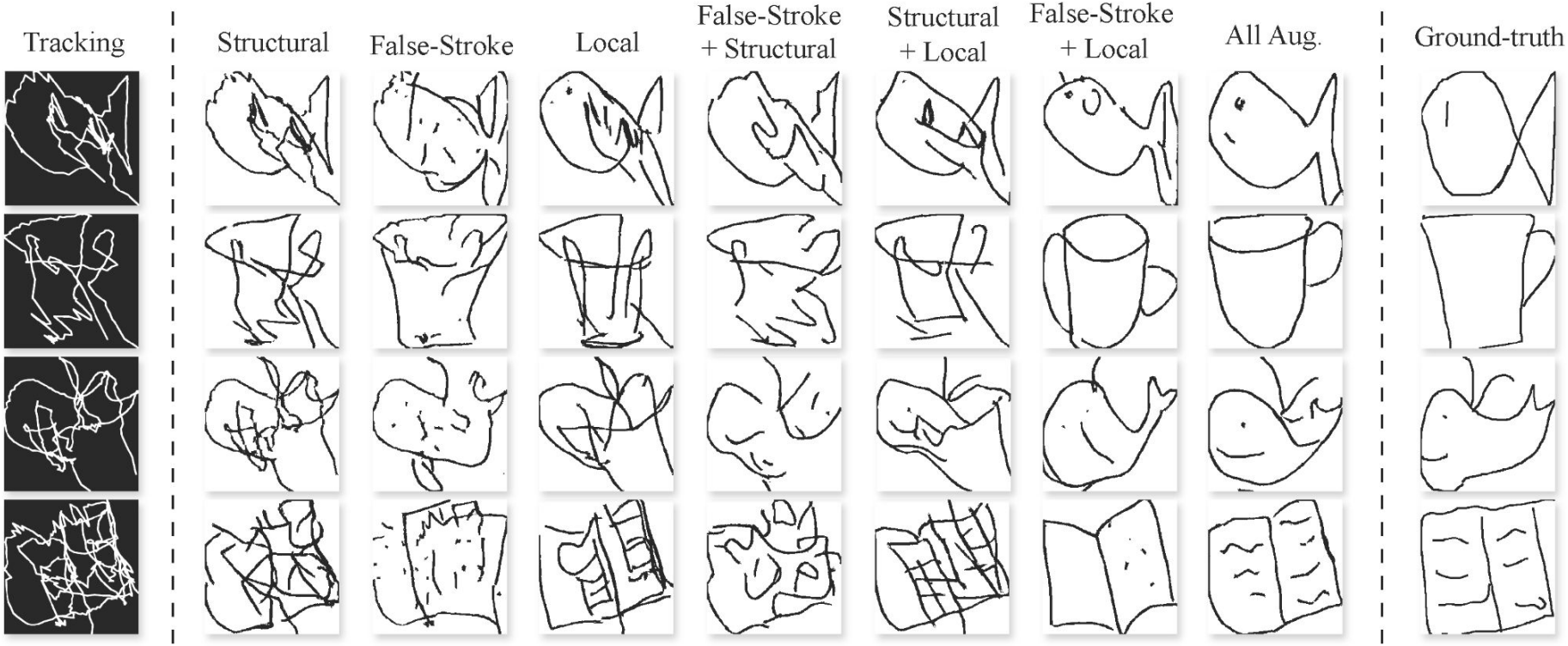
	Dataset	Backbone	w/ Aug.	SSIM (↑)	CD (↓)	LPIPS (↓)	CLIP I2I (↑)	CLIP I2T (↑)
Seen Categories								
Tracking	synth.	–	–	0.59	20.12	0.36	0.80	0.22
Gen.	synth.	SDXL	✗	0.59	20.11	0.37	0.79	0.23
Gen.	synth.	SD1.5	✓	0.60	17.98	0.35	0.80	0.26
Gen.	synth.	SDXL	✓	**0.64**	**17.39**	**0.33**	**0.85**	**0.28**
Tracking	real	–	–	0.55	32.36	0.42	0.76	0.21
Gen.	real	SDXL	✗	0.55	31.99	0.41	0.79	0.21
Gen.	real	SD1.5	✓	0.59	27.59	0.38	0.80	0.27
Gen.	real	SDXL	✓	**0.64**	**25.46**	**0.36**	**0.84**	**0.29**
Unseen Categories								
Tracking	synth.	–	–	0.59	20.47	0.36	0.80	0.22
Gen.	synth.	SDXL	✗	0.59	20.32	0.35	0.81	0.22
Gen.	synth.	SD1.5	✓	0.60	17.50	0.35	0.80	0.26
Gen.	synth.	SDXL	✓	**0.64**	**17.27**	**0.34**	**0.85**	**0.27**
Tracking	real	–	–	0.54	33.92	0.42	0.76	0.21
Gen.	real	SDXL	✗	0.55	33.53	0.41	0.78	0.21
Gen.	real	SD1.5	✓	0.61	27.67	0.38	0.80	0.27
Gen.	real	SDXL	✓	**0.63**	**24.26**	**0.38**	**0.85**	**0.28**



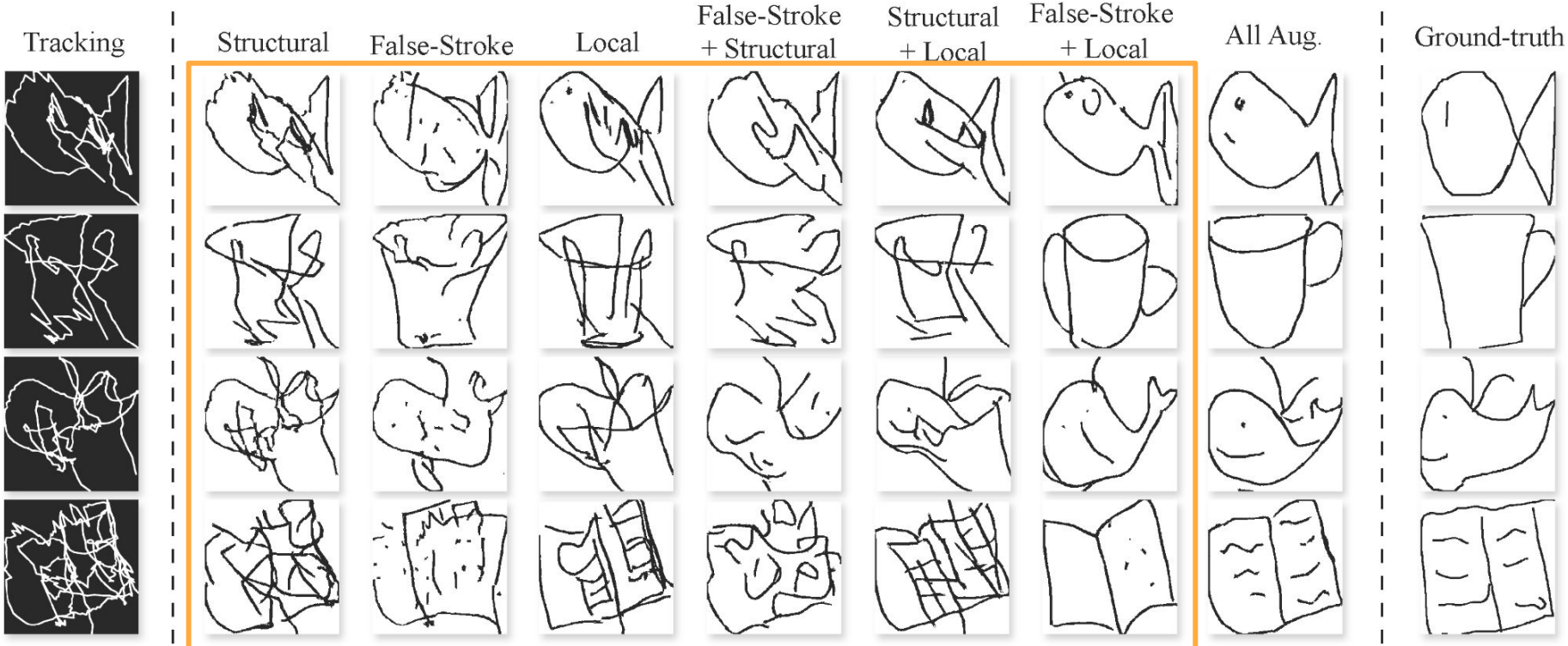
Generations on unseen categories from Quick, Draw! dataset.



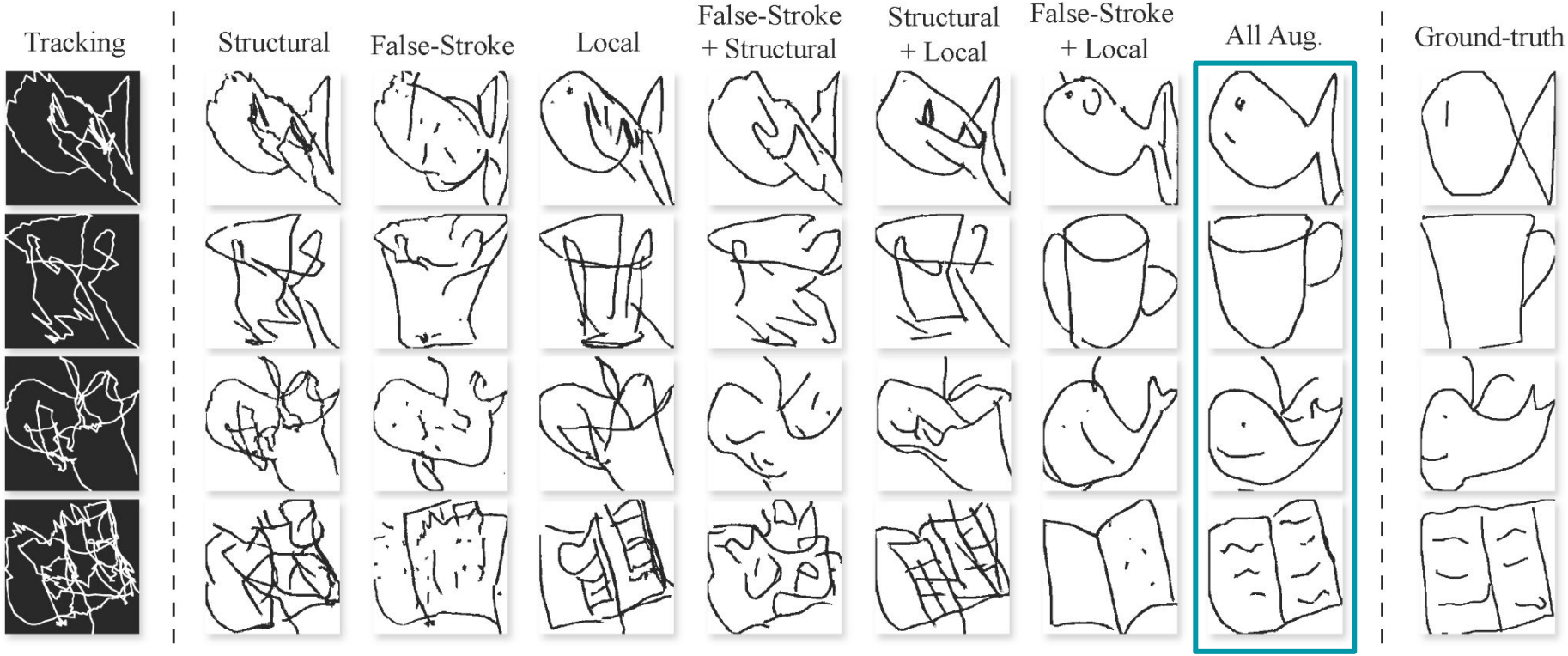
Generations on TUBerlin dataset



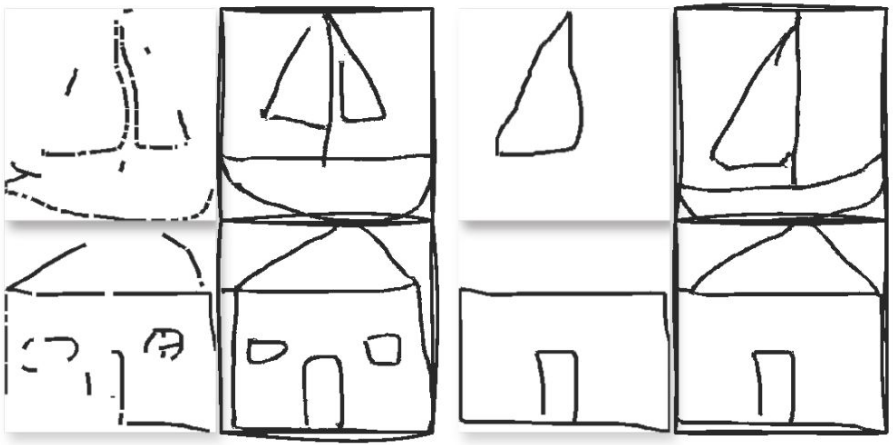
Generations on models with a subset of augmentations applied.

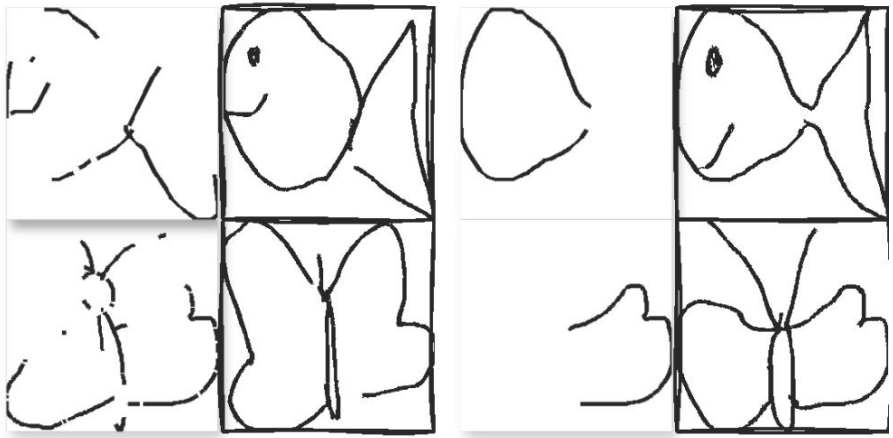


Generations on models with a subset of augmentations applied.

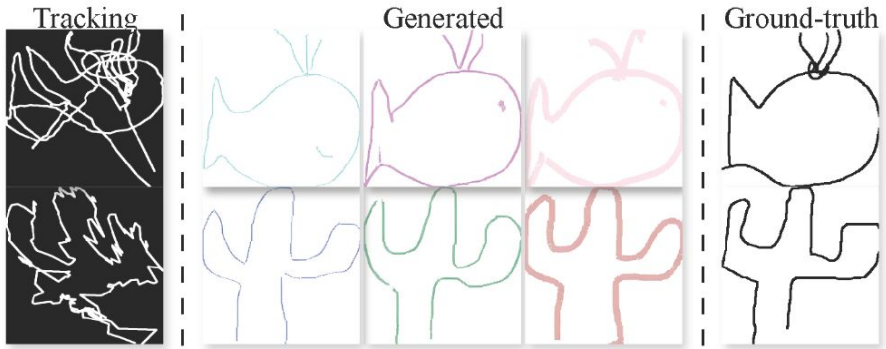


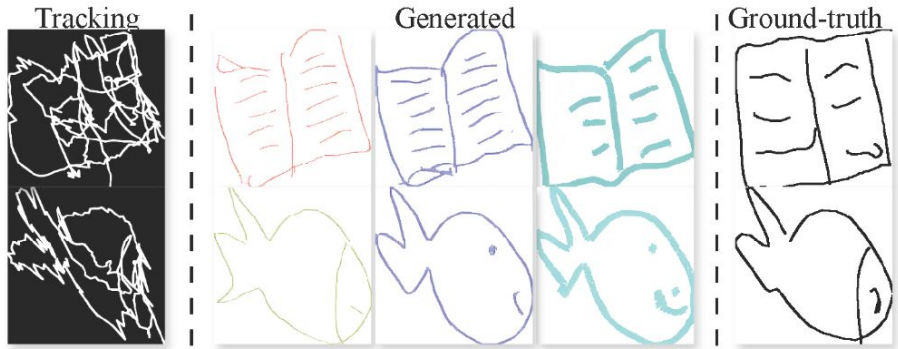
Generations on models with a subset of augmentations applied.





Sketch auto-completion for partially removed and wholly removed strokes





Generation with specific line specifications (color, thickness)

AirSketch: Generative Motion to Sketch

Hui Xian Grace Lim, Xuanming Cui, Yogesh S Rawat, Ser-Nam Lim