Reproduction and Reliability Evaluation of Segment This Thing for Efficient Point-Prompted Image Segmentation

Abstract

The official pretrained Model of a point-prompted foveated-tokenization based segmentation model STT-B (CVPR 2025) is evaluated via replication, applying the model external on Oxford-IIIT Pet with none retraining. 499 out of 500 samples have been validated successfully in the work. STT-B obtained 0.6070 IoU, 0.7149 Dice, 0.8550 precision, 0.7226 recall, and end-to-end latency of 93.0 ms. When comparing with the existing counterpart SAM-ViT-B on a paired 100-images subset, SAM yields higher accuracy (0.6717 IoU over 0.6306) but at much higher efficiency (93.0 ms over 440.0 ms). Further analyses including Prompt, degradation, calibration, geometry and failures demonstrate a favorable accuracy-efficiency tradeoff of STT-B; However, a poor boundary recovery performance and failures with high confidence have been discovered.

Keywords—*interactive segmentation, STT, SAM, foveated tokenization, robustness, calibration*

1. Introduction

Promptable segmentation aims to extract objects without requiring task-specific fine-tuning using prompts that are either points, boxes, or masks [2]. SAM is a powerful general purpose baseline [2], however large image encoders make frequent interaction costly. At CVPR 2025, Schmidt and New combe propose Segment This Thing (STT), an efficient alternative [1].

STT utilizes foveated tokenization where local points of high spatial resolution near the prompt's position are maintained, whilst areas away from the prompt are modeled more coarsely. This leads to an efficient set of tokens and offers interactive rate inference, with reduction in size not directly achieved by simply scaling down the model[1].

We do not attempt any benchmark reproductions or try to achieve the state-of-the-art. The motivation is instead how much quality the system STT-B preserves over increased interaction rate. We also provide analysis on where the release fails when evaluated on held-out data.

2. Related Work

SAM is an efficient promptable foundation segmentation model for zero-shot transfer [2]. The ViT-B Checkpoint provided can serve as a straightforward baseline for STT-B. The external distribution for evaluation are categories of Oxford-IIIT Pet which has 37 different classes with different poses, lighting condition, sizes etc and provides pixel level trimaps [3].

The efficient inference of STT is achieved through input foveation, meaning the crop is always in the neighborhood of the click with pixel resolution linearly decaying as you zoom out [1]. The click serves to provide information on the salient regions in which finer pixel resolution is needed.

Study Contributions

- Reproducing our official STT-B inference pathway on the released checkpoint and predictor. A 499-sample external segmentation evaluation with bootstrap 95% confidence intervals. A paired STT-B vs SAM ViT-B on the identical 100-image prompt protocol.
- Ablations on point positioning and image corruption, as well as a confidence and catastrophe analysis.
- Object-geometry errors beyond standard mask metrics.

3. Problem Definition

We use the official pretrained STT-B model to predict a binary segmentation mask M for an input RGB image I and a positive point prompt p. No additional training or fine-tuning is performed. Instead, we reproduce the released inference pipeline and evaluate the pretrained model on an external dataset to study segmentation quality, efficiency, robustness, reliability, and failure behaviour.. We test for overlap with Oxford ground truth, sensitivity to p and image degradation, speed against SAM, and whether severe failure instances are captured by the STT score.

4. Methodology

4.1 STT-B inference

We utilize the publicly released STT-B pre trained weights and predictor, without any training or fine-tuning. For a given image, we provide STT with the RGB input along with one positive point. For the candidate(s) returned by STT, we select the one whose predicted-IoU is the largest.

The generated foveated output is then reconstructed using the released pattern, followed by sigmoiding, thresholding at 0.5, and re-mapping the coordinates to the image space with the prompt-centered receptive-field crop.

4.2 Dataset and foreground convention

Dataset of 3,669 test images for Oxford-IIIT Pet [3]. Label 1 is foreground, label 2 and 3 is non-foreground. The image with less than 100 foreground pixels is discarded.

One sample was skipping from 500 images, so we selected 499.

N = 499.

4.3 Small-image preprocessing

Scaling if shortest dimension of an image is below 1280 (then expand to 1280, scaling the point also by 1280 / original length of that side). Resize back the binary prediction by nearest neighbors and calculate the score. Mean input scale is 3.79.

4.4 Primary point prompt

This deeply interior, foreground pixel obtained by the distance transform, which is the one to be used according to the pre specified main prompt: points that are centroid, inside interior of image, or near boundary are all simply ablated and are not inserted back post-hoc into the 499 image result.

5. Evaluation Protocol

5.1 Segmentation metrics

We output IoU, Dice, precision, recall, and Boundary F1. IoU and Dice measure the amount of overlap; precision is inversely related to the amount of pixels in P that lie outside G, and recall is inversely related to the amount of pixels in G that lie outside P. IoU= |P G| / |P U G| , where P and G are the predicted and ground-truth segmentations, respectively, and U and represent set union and intersection.

Dice= 2|P G| / (|P| + |G|) .

Boundary F1, which accounts for prediction error near boundary contours, was calculated at 2 pixel boundary tolerance at the original image resolution and should be assessed distinctly because of the potential to disproportionately alter the contours due to tri map to binary segmentation mapping and scaling.

5.2 Efficiency metrics

Model Runtime is runtime of STT predictor; End-to-End Runtime adds scaling, reconstruction and remap to the original resolution. CUDA is synchornized before and after measured operations, and FPS = 1 / end-to-end latency, and highest allocation on the GPU is recorded.

5.3 Final experiment design

The main experiment needs 500 seed-42 test samples, 499 success results come from it. For Prompt/degration experiment, it relies on the same constant subset (100 images). SAM chooses 100 main-set images and the same deepest-interior point.

5.4 Statistical analysis

Bootstrap Resampling is used to perform 95% confidence intervals on the primary-metric. Catastrophic failures are defined as any IoU < 0.25, whereas confident failures also demand a predicted IoU > 0.70. We investigate Predicted versus Actual IoUs using Pearson correlation and a linear fit.

5.5 Implementation environment

All experiments conducted within Google Colab, with hardware of NVIDIA Tesla T4. Corrected end-to-end timing measure includes operations involved in practical usage of; hence, SAM image embedding is taken into account in the baseline latency measure.

6. Main Results

The STT-B final 499-sample result is summarized in table 1. The overlap between regions are reasonable. Moreover the precision performance are almost stronger than the recall.

That the reason is that the prediction foreground is usually useful but model under-segmenting object, i.e.

Omitting a part of object.

Table 1. Final STT-B segmentation and efficiency results (n=499).

Metric	Mean
IoU	0.6070
Dice	0.7149
Precision	0.8550
Recall	0.7226
Boundary F1	0.0952
Model time	0.0283 s
End-to-end	0.0930 s
End-to-end FPS	11.00
Peak GPU alloc.	0.397 GB
Area error	34.85%
Perimeter error	27.06%

The intervals from table 2 have been found to be significantly narrower than the previously previously estimated for the earlier, 20 images sanity run. This illustrates what a more definitive final sample must be to report values from.

Table 2. Bootstrap 95% confidence intervals.

Metric	Mean	95% CI
IoU	0.6070	0.5844–0.6280
Dice	0.7149	0.6918–0.7362
Precision	0.8550	0.8464–0.8641
Recall	0.7226	0.6943–0.7487
Boundary F1	0.0952	0.0881–0.1024
E2E time	0.0930	0.0917–0.0943

6. Interpretation of the Primary Evaluation

6.1 Accuracy characteristics

The values of MeanIoU 0.607 and Dice 0.715 show useful but not state-the-art region overlap on an external dataset. The highest conventionally accepted metric is precision 0.855. The recall is lower 0.723 thus the main mode of error consists of neglecting pixels of the object, not including lots of unrelated foreground in addition.

The values for boundary F1 are significantly lower, at 0.095.

We must not conceal this, but our result is an evidence that quite acceptable localization is obtained even when there is not very precise alignment of boundaries. This low score of boundary is also a logical consequence of the geometry expansion, on which the perimeter error is 27.06% and the mean area error is 34.85 %.

6.2 Efficiency

Raw STT predictor times average 28.3 milli second or around 35 model-only inferences per second. Adding the image resizing, reconstruction, and remapping back to the original image bumps the mean latency to 93.0ms or about 11 end-to-end FPS. The latter is a more relevant number for this reporting.

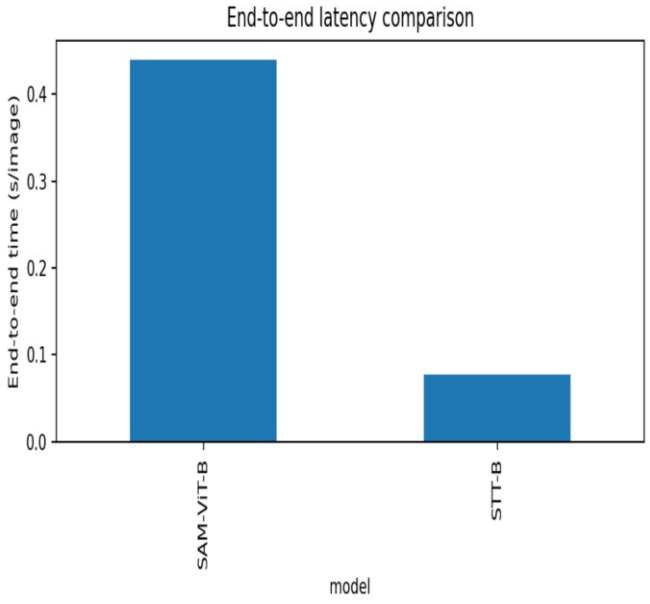
6.3 Why the final IoU is lower than the pilot

A previous 40-image pilot achieved a mean IoU higher than 0.65, but was based on a much smaller dataset and previous preprocessing settings. The 499-image experiment seems far more robust, since it has small image preprocessing corrected, fixed large image prompt, and almost five hundred times more data. A large experiment could also reveal low-frequency catastrophic failure that a small-image experiment would almost certainly miss.

6.4 External-evaluation scope

Don't compare these numbers directly with CVPR benchmarks. The original STT paper considered another suite of benchmarks and the configuration/hardware [1]. We wanted to ask answer how the publicly released STT-B model behaves on Oxford-IIIT Pet with the clear external procedure.

7. STT-B versus SAM ViT-B



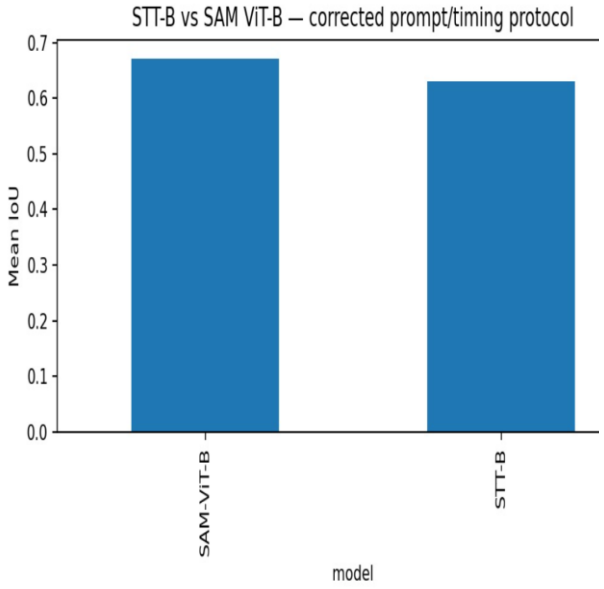


Fig. 1. Accuracy–efficiency trade-off on the paired 100-image baseline. Insert the IoU chart on the left and runtime chart on the right.

We then test SAM ViT-B on these same 100 images, and use the same deepest-interior positive point. In particular, SAM image embedding is timed as part of its run, so the resulting number corresponds to a single-image interactive experience, not just the mask-decoder time alone.

Table 3. Paired STT-B versus SAM ViT-B comparison (n=100).

Metric	STT-B	SAM-B
IoU	0.6306	0.6717
Dice	0.7368	0.7686
Precision	0.8472	0.7952
Recall	0.7589	0.8685
Boundary F1	0.1007	0.0636
Runtime	0.0768 s	0.4400 s
FPS	13.54	2.28

SAM gains a 0.0411 improvement for mean IoU and 0.0318 improvement for Dice. Additionally, it achieves higher recalls which mean that it covers objects with better completeness. STT is higher on precision and higher Boundary F1 under this implementation.

7.1 Accuracy–efficiency trade-off

The main point isn't that STT somehow wins big in overall average overlap. It wins by about the same amount on average as STT does. The difference is on speed: STT is very much fast-2.53 vs 0.30 seconds per image on average; 5.7x on speed.

STT has 13.54 average FPS whereas SAM gets 2.28.

(That's an advantage of almost 6-timesmorethroughput for STT). Because the comparison is paired upon every image, we also see variety: STT wins the overall average overlap on 46 of 100 images, while SAM takes home the win for the other 54 images.

7.2 Statistical interpretation

The pair wise IoU difference, STT minus SAM, is 0.0819 to 0.0017 at 95% CI. The CI implies that SAM tends to be better for overlap while the timing difference seems certain in the reverse. Given a latency sensitive interactive application, it implies for us to plot a quality versus speed boundary.

7.3 Practical implication

In effect, STT-B's better description is the most performant approach rather than a universal superiority over SAM. To STT, it's acceptable to have an unacceptably small mean overlap so long as the task is performed quickly enough. Where completeness is more important than speed, however, paired result is clearly for SAM.

8. Prompt Robustness

This work is a case of point prompted segmentation and it requires user's clicks location. We compare 4 different deterministic/controlled prompt methods using the same 100 images.

Table 4. Prompt-location robustness on the paired 100-image subset.

Prompt	IoU	Precision
Centroid	0.6498	0.8116
Deepest interior	0.6161	0.8465
Random interior	0.5318	0.8212
Near boundary	0.5317	0.7866

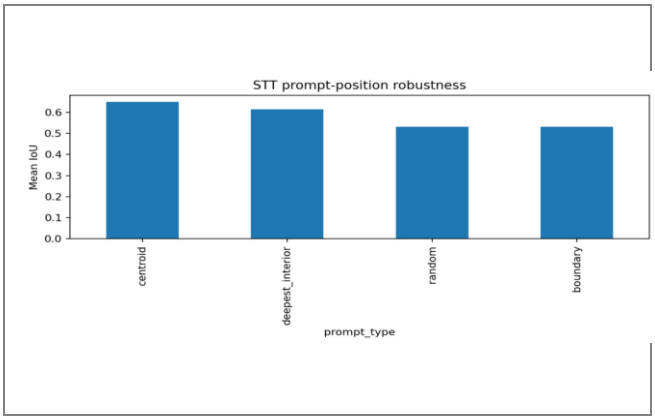


Fig. 2. Mean IoU under four point-placement strategies.

The centroid generated the top mean IoU (0.6498) and the top recall (0.7818), and the deepest interior has the top precision (0.8465). The random-interior and near-boundary clicks were of significantly less quality. This indicates that prompt dependency was genuine.

The centroid result was the result of ablation and was not used post-hoc to change the predefined primary prompt type. It would not have been the case if the deepest-interior protocol were used with all 499images evaluated.

9. Image-Degradation Robustness

We perform the same robustness on the subset of 100 images under: blur, a short round trip to low-res and back, dark, bright. We keep the point location fixed, so transformations more to the image instead of the prompt itself

Table 5. Degradation robustness (same 100 images).

Condition	IoU	Dice
Original	0.6161	0.7267
Blur	0.6004	0.7099
Low resolution	0.6079	0.7189
Dark	0.6286	0.7372
Bright	0.5744	0.6853

9.1 Observed behavior

The blurring effect decreases the IoU by roughly 0.016 and the low-res by approximately 0.008 when compared to original subset. The brightening effect causes the biggest negative drop in IoU of approximately -0.042. It is interesting to note that the darkened case has a slight positive effect and the average IoU actually rose to 0.6286. We interpret this not as any evidence that darker scenes are typically easier, but rather as a predictable if non-monotonic response to particular conditions.

9.2 Why no extra degradation figure

Table 5 effectively shows the ranking and size of degradation for an eight page paper. Figure budget should be saved for analyses which cannot be interpreted as readily from a table such as the baseline trade-off, prompt sensitivity, calibration and qualitative mask behavior.

10. Confidence Calibration and Failures

We are generating a candidate mask and STT produces a predicted-IoU value between them. We then calculate the ground truth IoU and compare this with STT's prediction. Pearson correlation over 499 samples is 0.518. This suggests a moderately positive relationship, with quite a lot of spread.

Table 6. Confidence and catastrophic-failure summary.

Statistic	Value
Samples	499
Predicted/actual IoU corr.	0.5182
Linear slope	1.0651
Linear intercept	−0.2318
Actual IoU < 0.25	66
Confident failures	22
Failure rate	13.2%
Confident-failure rate	4.4%

10.1 Failure definitions

We defined a catastrophic failure to be any real IoU<0.25. Out of 499 total images, 66 pass this condition (13.2% of all judged samples). A confident failure is any that additionally predicts IoU>=0.70. There are 22 such cases.

The most salient cases are those 22 that illustrate that model internal score is not an indicator that a predicted mask is reliable. About one third of all catastrophic failures are confident cases using this threshold:

10.2 Reliability interpretation

The use of only the average metrics would mask such a failure mode. Instead, the calibration analysis broadens the effort from a pure performance measurement to one also including reliability estimation, because a system could appear fast and moderately accurate on average but still make one or a few extreme mistakes for which none of the confidence scores were particularly elevated.

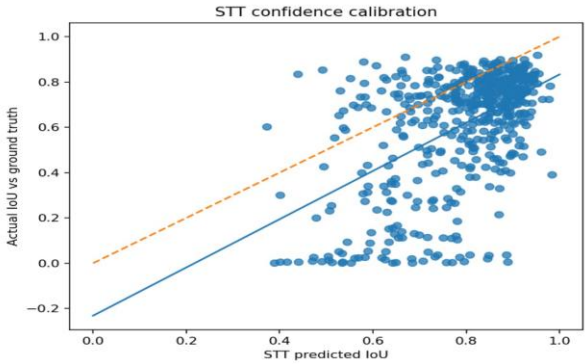


Fig. 3. STT predicted IoU versus actual IoU. The diagonal indicates perfect calibration.

10.3 What the calibration plot shows

There is an overall trend for higher actual IoU to be produced as the predicted IoU increases, reflected in the 0.518 correlation coefficient. However, there is considerable spread of results: quite a few are much lower than the perfect diagonal, including some with high predicted IoU and a very low actual overlap. The fitted line only reflects this average behavior, not that each image should lie on it.

11. Object-Geometry Extension

The notebook also computes foreground area, perimeter, bounding box, occupancy, and relative area/perimeter error. Mean area error is 34.85% and mean perimeter error is 27.06%. These large geometry errors are consistent with the low Boundary F1 and support the same conclusion: STT often can localise the right object but might miss or oversize the accurate shape.

The geometry extension is particularly handy in terms of interpreting downstream suitability. Rough interactive extraction, object picking, or annotation support might accept imprecise contour; metrology, medical boundaries, and precision manufacturing would need more robust shape recuperation and validation.

12. Qualitative Analysis and Discussion

Fig. 4. Required qualitative evidence: original image and prompt, ground truth, STT prediction, and overlay. Include both strong segmentations and at least one catastrophic failure.

Refer the image in the last page .

12.1 Visual evidence

The qualitative panel is understandable as you cannot infer from the region metrics how spatial your errors are even if you get region metrics. You get two reasonably helpful segmentations in the final output picture in the range of around IoUs from 0.73-0.85 along with one failure in the range of around IoU 0.024 wherein almost none of the object is shown. And you find a pair of successful results where you need only one pixel to get the full dog/cat, however, with varying backgrounds and postures.

Also you do see some failures where the point you point at falls inside the correct region, but you would get only a minuscule local reconstructed area.

It seems like better to show two outputs, the region metrics and the qualitative panel, so that readers can be shown successful outputs and not just "how failed the segmentation fails visually".

12.2 Overall discussion

All experiments together provide further evidence for an accuracy-efficiency reading of results. While STT-B doesn’t win on accuracy amongst the paired baselines, its latency is much lower. Good on precision, OK on moderate blurring and low-resolution variations, quality definitely hinges on the position of the prompt. The main failure mode is on the score of robustness: boundary quality is low and it presents 66 catastrophic failures in the 499 image dataset.

This may suit it to interactive applications where the user gets the immediate quality of the mask and can interactively correct it, but perhaps not as an automatic mask generation tool when a single error could have high costs.

13. Limitations

1. It is an external dataset and not the entire set of benchmark images from the STT paper [1]. It measures the tendency towards transferred behavior instead of exact benchmark repetition.
2. It does not include a fine-tuning stage.

Using domain-specific training could influence accuracy and confidence behavior as well.

3. The annotation in Oxford is a trimap.
The binary conversion used to turn the provided trimaps to binary mask classifies label 1 as fg and 2-3 as bg. 4. Heavy oversampling is performed (mean scale of 3.79) prior to STT inference. The interpolation may destroy fine edges.
5. The single positive point is used in the main experiment and several of the failures are correctable with other positive and negative points.
6. The SAM benchmark has only 100 positive images instead of the main 499 images. 7. The reported boundary F1 is dependent on the binary decision for pixels two pixels away.

14. Future Work

The entire 3,669-image Oxford test split along with a handful of outside collections. We consider STT-B vs higher-parameter STT models, and several competitive segmentation baselines. We experiment on multi-click correction and negative queries post first failure. We either calibrate predicted IoU or train a light failure predictor to detect unlikely predictions.

15. Reproducibility and Practical Implications

15.1 Reproducibility

Notebook 14 fixes model experiment sizes and seeds, loads the official STT-B checkpoint, runs experiments on SAM in three configuration modes: the main SAM experiment, two robustness versions, and the paired SAM experiment, and saves CSV/JSON summaries and output plots.

Main, robustness and baseline experiment sizes are requested as 500, 100 and 100 respectively, whereas 499 main runs were recorded in the report. Both model only and end to end runtime of individual steps is also logged separately. Baseline measurement in the report usescorrected end-to-end time including SAM embedding.

15.2 Practical implications

What works? In a laboratory / prototype setting, the project demonstrates that a new CVPR model can be cloned, retrained, and tested against accuracy, speed, reliability, prompt sensitivity, and robustness. From a practical standpoint, STT-B seems like a good choice if speed (latency) is the most important criterion and a human observer can check the resulting mask – if thoroughness and reliability are the priority then some form of a higher certainty criterion is likely necessary

16. Conclusion

We replicated STT-Band measured it against external data using Oxford-IIIT Pet. Over 499 successful images, STT-B attained a measured IoU of 0.6070, a Dice of 0.7149, precision 0.8550, and recall 0.7226. Measured with 100pairedimages with SAM to assist the process it gained an extra 0.0410IoUwhile it gained 5.7x more latency.

Placing prompts affects precision fairly significantly.

However, our alterations to images (low resolution or blur) seemed to have a less damaging effect than our adjustments increasing the image luminosity, although overall placing prompts is significant to improving accuracy. A predicted IoU of the actual IoU had a correlation of0.518, and we found22catastrophic failures at high confidence. Our methods provide a useful efficiency accuracy trade-off but boundary quality could be of greater consistency and reliability.

References

[1] T. Schmidt and R. Newcombe, “Segment This Thing: Foveated Tokenization for Efficient Point-Prompted Segmentation,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 29428–29437, 2025.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023.

[3] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, “Cats and Dogs,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2012.

[4] Facebook Research, “Segment This Thing: Official Implementation and Pretrained Checkpoints,” GitHub repository, 2025.

[5] Meta AI, “Segment Anything: Official Implementation,” GitHub repository, 2023.

[6] To optimize manuscript clarity and explore conceptual frameworks, generative artificial intelligence tools—specifically ChatGPT and Claude—were utilized during the drafting phase for sentence rephrasing, grammatical editing, and exploratory brainstorming and code optimization.

GitHub_Repository:
https://github.com/gundeeps247/cv_dl

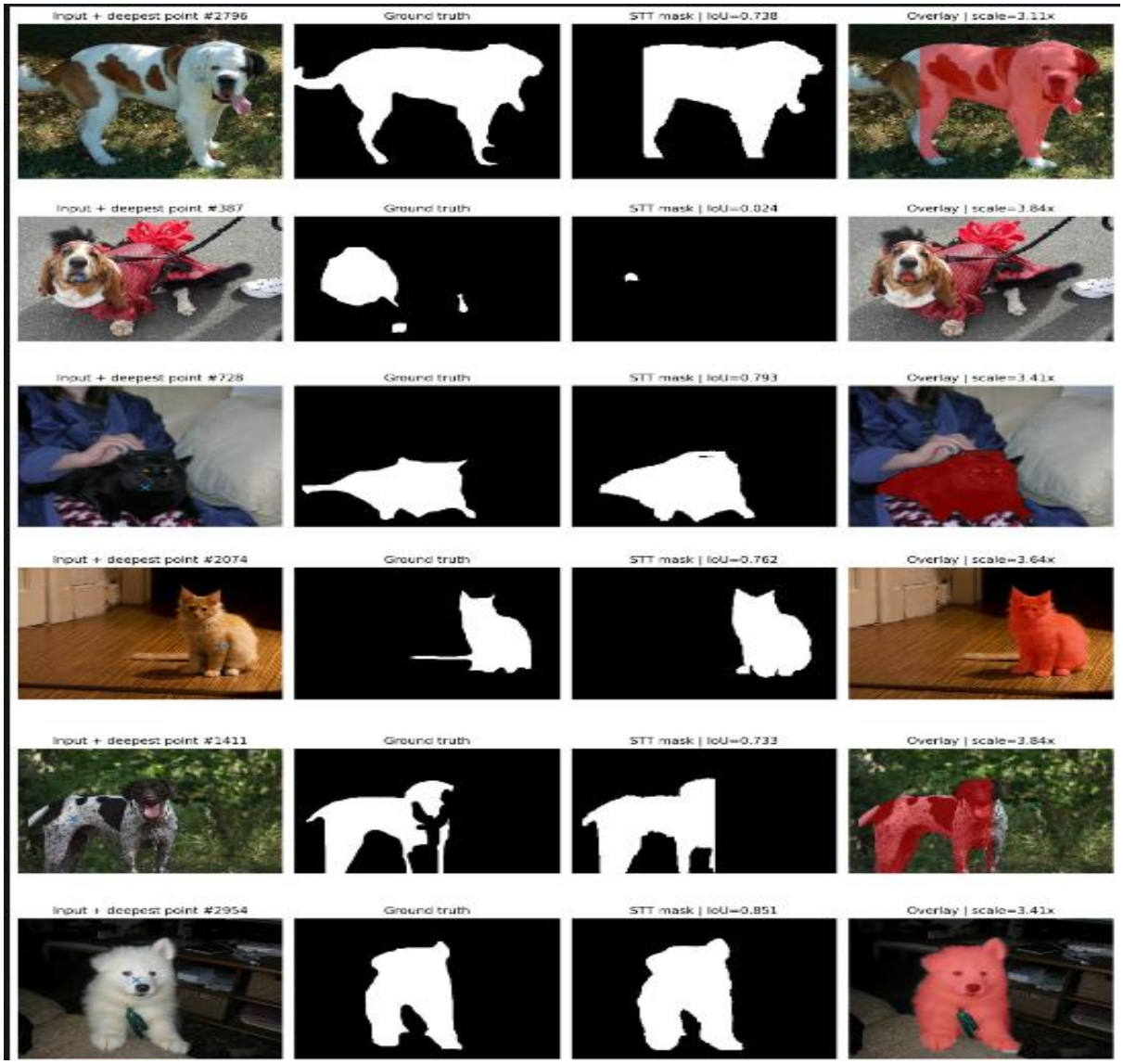


Fig. 4. Required qualitative evidence: original image and prompt, ground truth, STT prediction, and overlay. Include both strong segmentations and at least one catastrophic failure.