




ARTICLE

Supplementary Materials: Multiple Point MedSAM Prompting for Enhanced Medical Image Segmentation

Wasfieh Nazzal¹, Ezequiel López-Rubio^{1,2,3} , Miguel A. Molina-Cabello^{1,2,3}  and Karl Thurnhofer-Hemsi^{1,2,3,*} 

¹Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Bulevar Louis Pasteur, 35, Málaga, Spain

²ITIS Software, Universidad de Málaga, C/ Arquitecto Francisco Peñalosa 18, Málaga, Spain

³Biomedic Research Institute of Málaga, IBIMA Plataforma BIONAND, C/ Doctor Miguel Díaz Recio, 28, Málaga, Spain

*Corresponding Author: Karl Thurnhofer-Hemsi. Email: karlkhader@lcc.uma.es

Received: 11 December 2025; Accepted: 26 January 2026

S1 Datasets

A detailed summary of the nine datasets used to evaluate the proposed method is presented in Table S1. This table details the modality, the number of segmentation labels, the number of processed patient scans, and the total number of 2D testing slices used for inference.

Table S1: Summary of the nine benchmark datasets used in this study.

Dataset	Modality	Labels	Processed Scans	Testing Slices (2D)
FLARE2022_Abdominal [1]	CT	12	50	21187
MSD_Spleen [2]	CT	1	41	1052
MSD_Liver [2]	CT	2	131	21875
MSD_Pancreas [2]	CT	2	281	9487
MSD_Colon [2]	CT	1	123	1390
KiTS_Kidney [3]	CT	3	283	56380
MSD_BrainTumor [2]	MRI	3	484	36952
MsLesSeg_MSLEsion [4]	MRI	1	52	2654
LASC_Heart [5]	MRI	1	20	1337

S2 Parameter Analysis: Boundary Adherence (NSD)

Fig. S1 presents the Normalized Surface Dice (NSD) results for the hyperparameter analysis on the FLARE2022 dataset. These plots complement the DSC analysis presented in the main manuscript, highlighting the method's ability to improve boundary adherence, particularly for smaller objects.

S3 Experiment 1: Semi-Automatic GT-guided Prompt

This paper presents quantitative findings of Experiment 1. It describes the results of the aggregation strategies on the validation datasets. Note that the Mean Aggregation analysis of the

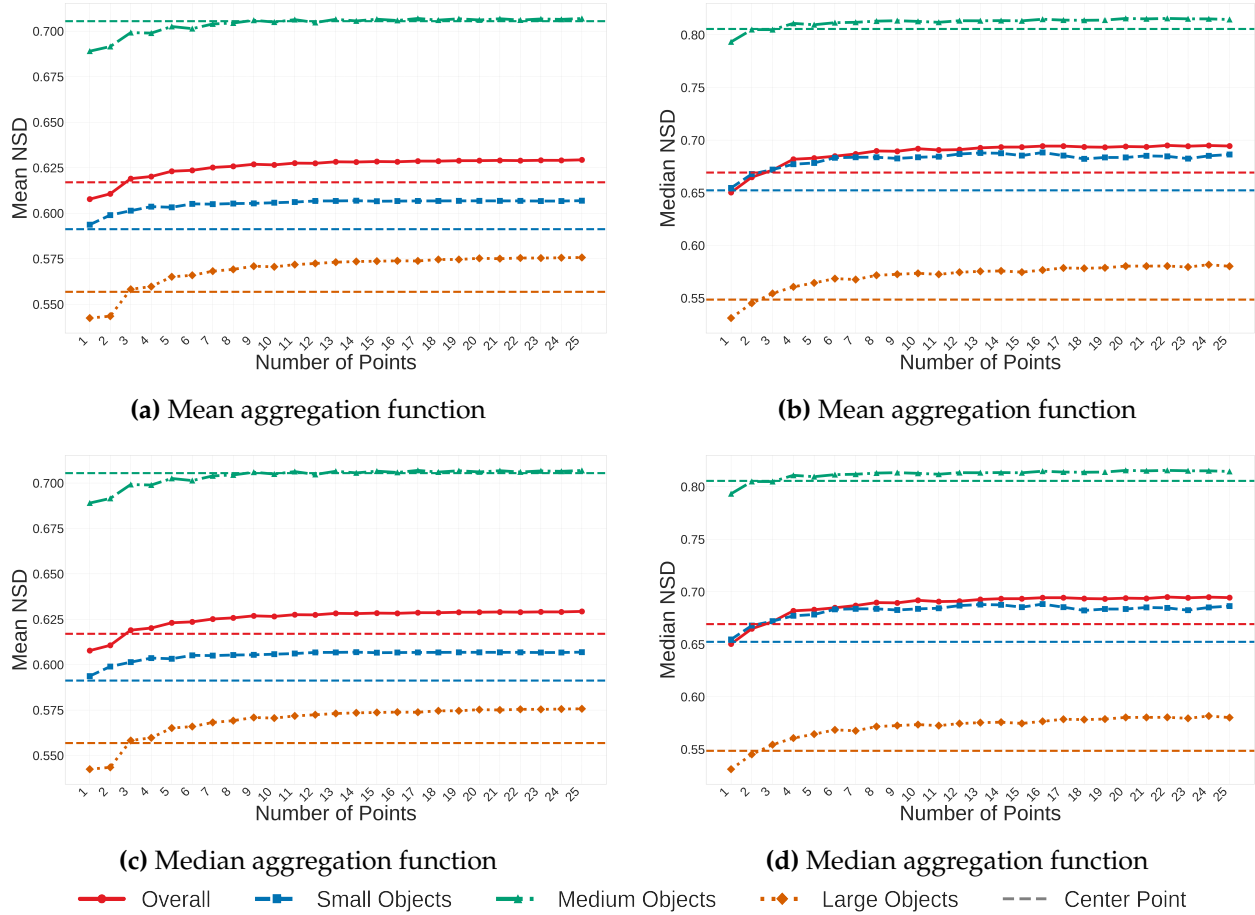


Figure S1: Normalized Surface Dice (NSD) parameter analysis on FLARE2022. Top row: Mean Aggregation; Bottom row: Median Aggregation.

KiTS data of the KiTS_Kidney is reported in the main manuscript; results of the other 7 datasets are given below. The Median Aggregation analysis includes all eight datasets, namely MSD_Spleen, MSD_Liver, MSD_Pancreas, MSD_Colon, KiTS_Kidney, MSD_BrainTumor, MsLesSeg_MSLesion, and LASC_Heart.

S3.1 Mean Aggregation Method

The following Figs. S2–S9 reveal the performance patterns of the Mean Aggregation strategy. Both figures give the Mean and the Median Dice Similarity Coefficient (DSC) and the Normalized Surface Dice (NSD) of prompt points (N) between 1 and 14.

The findings on the seven datasets that constitute this supplementary material exhibit a consistent log growth trend. In anatomically distinct structures (including the spleen, liver, etc.), segmentation accuracy quickly becomes stable and in many cases, scores near optimal DSC with as few as 3 or 5 points. Conversely, complex tasks with irregular boundaries, like those in colon and brain tumor segmentation, have a slower speed of performance improvement, which implies that a greater number of prompts is required to clear the ambiguity of the boundaries. The marginal improvement in accuracy reduces dramatically when past 10 points, and the convergence is usually

achieved at $N = 14$, which is consistent with this value as a good balance point of the main experiments.

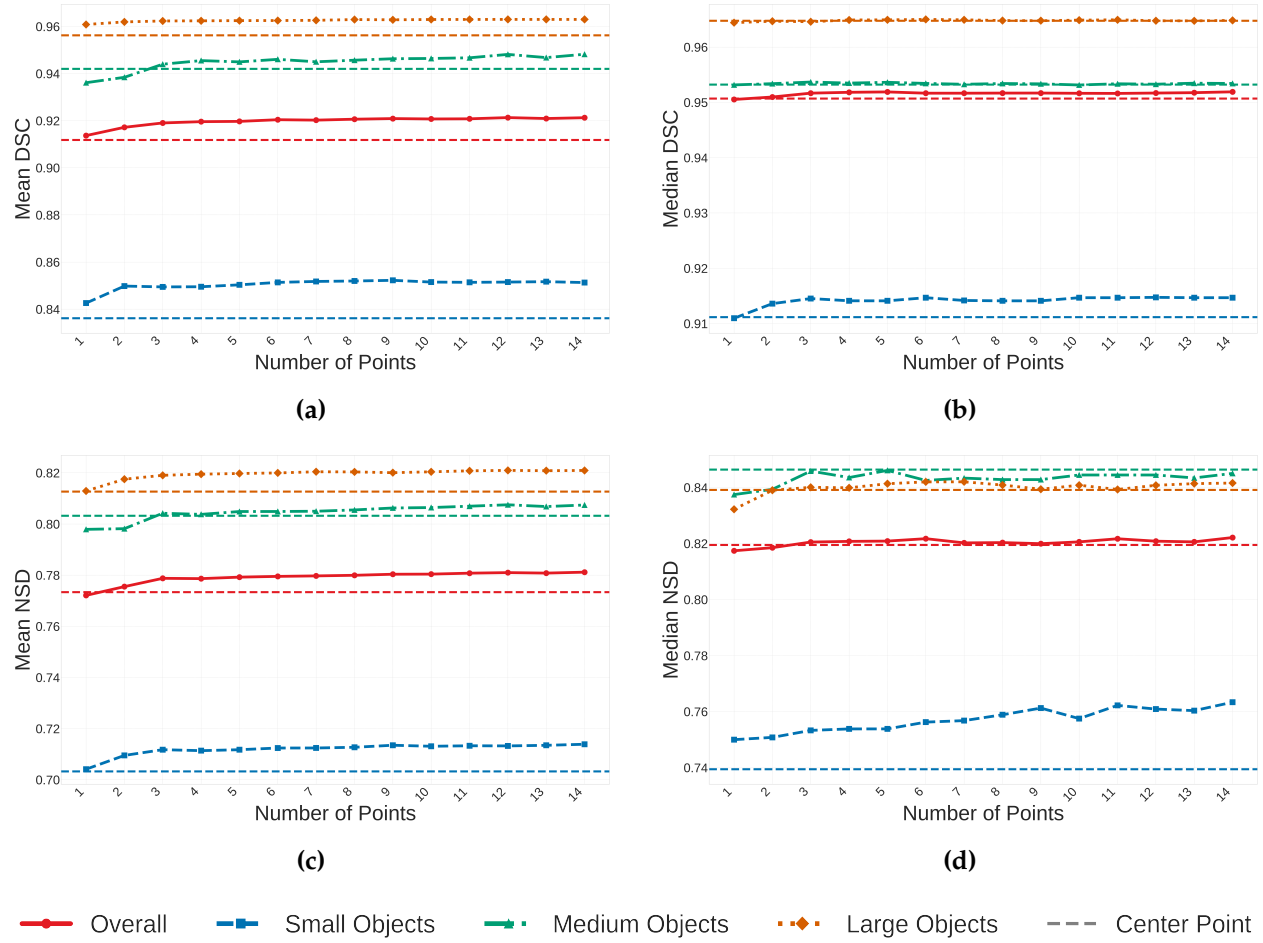


Figure S2: Mean Aggregation analysis for MSD_Spleen. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

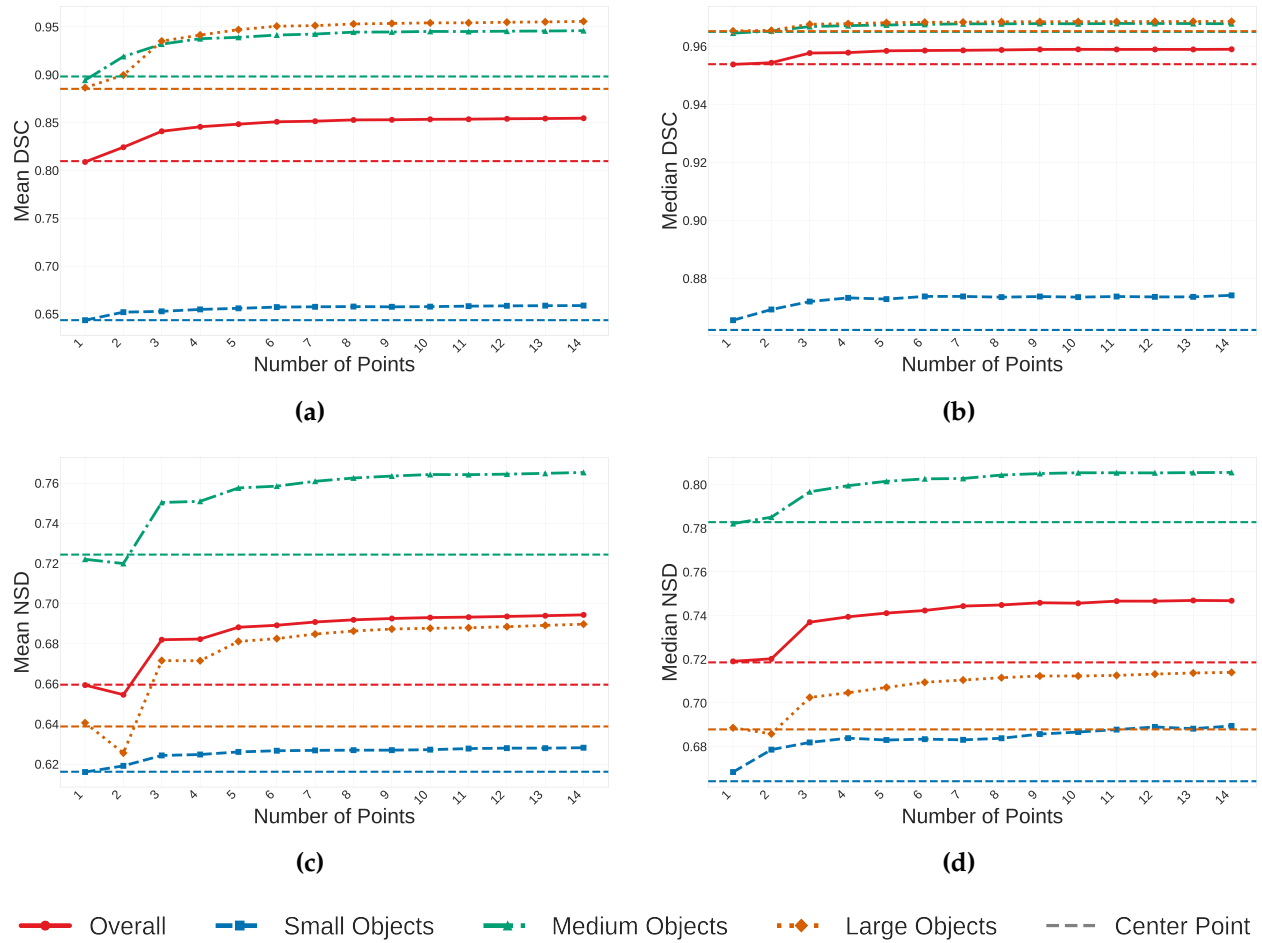


Figure S3: Mean Aggregation analysis for MSD_Liver. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

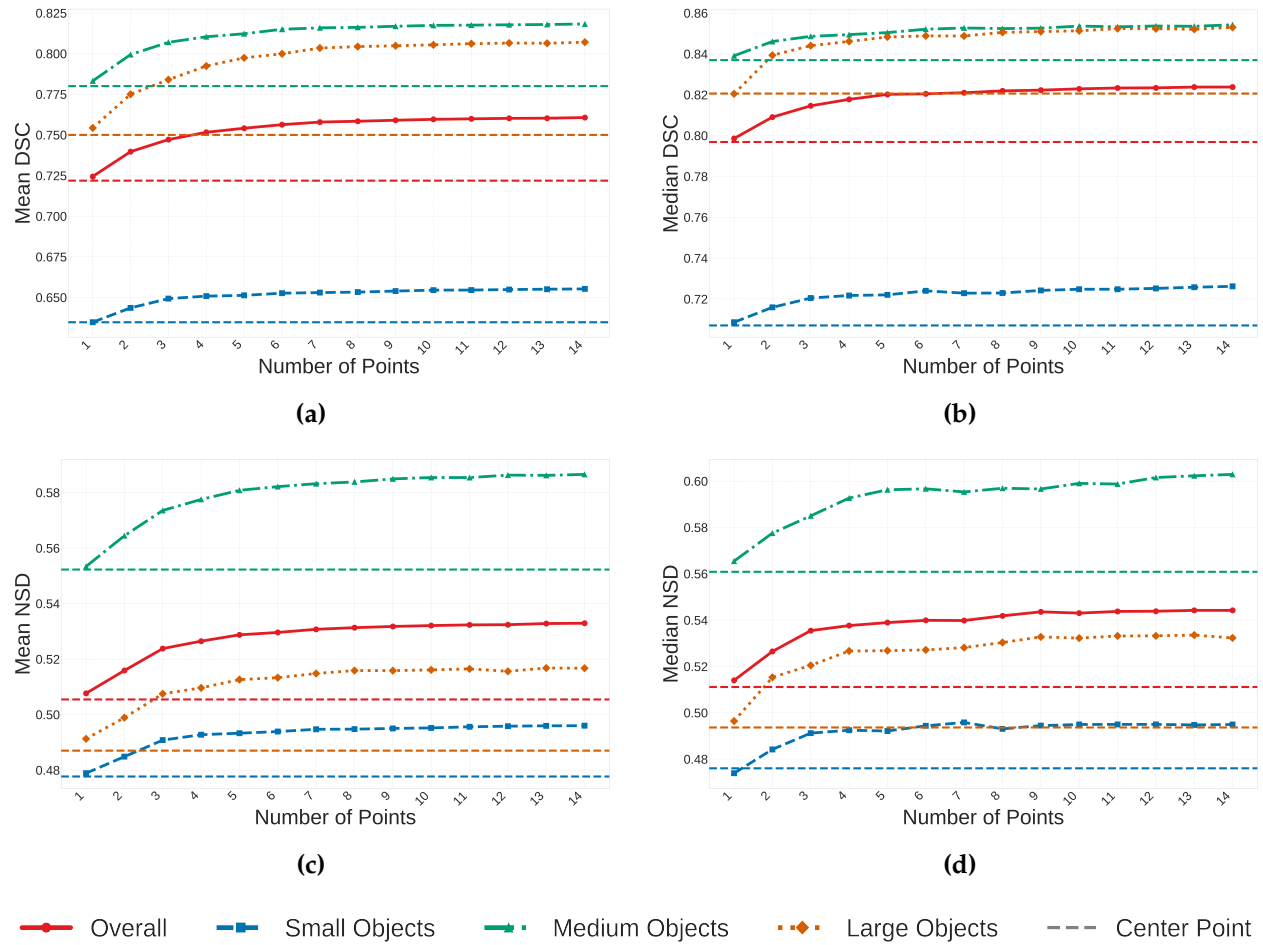


Figure S4: Mean Aggregation analysis for MSD_Pancreas. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

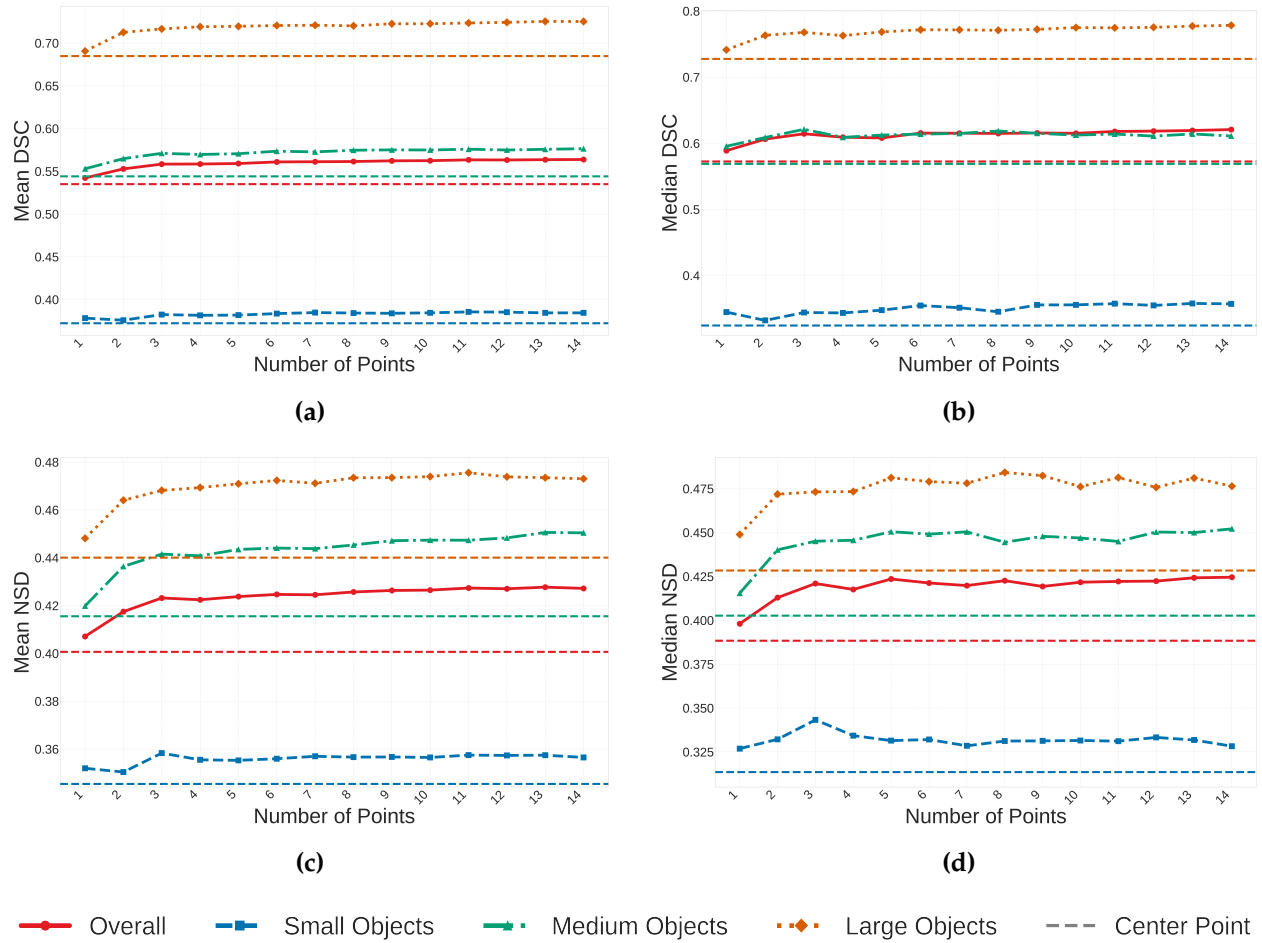


Figure S5: Mean Aggregation analysis for MSD_Colon. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

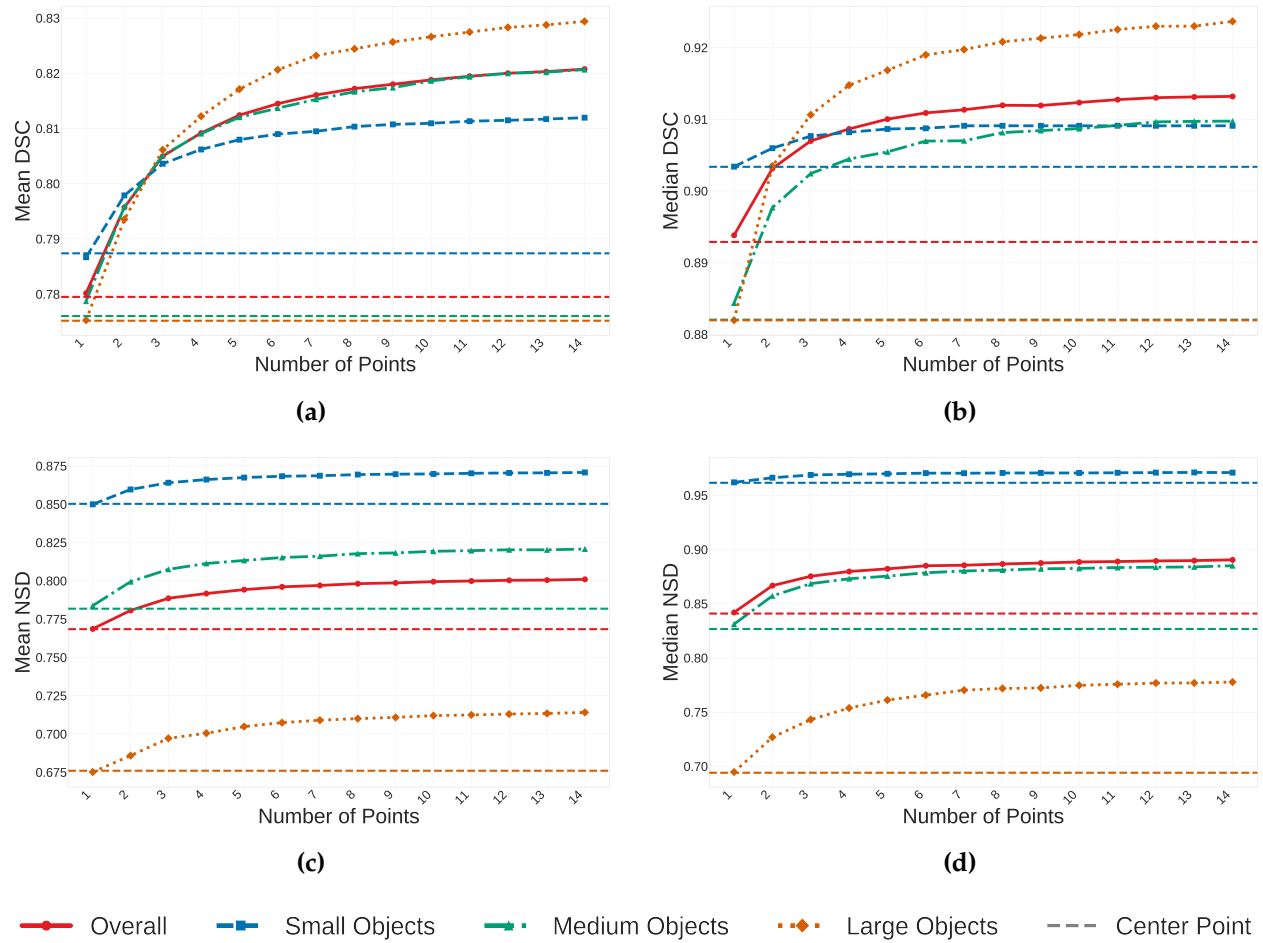


Figure S6: Mean Aggregation analysis for KiTS_Kidney. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

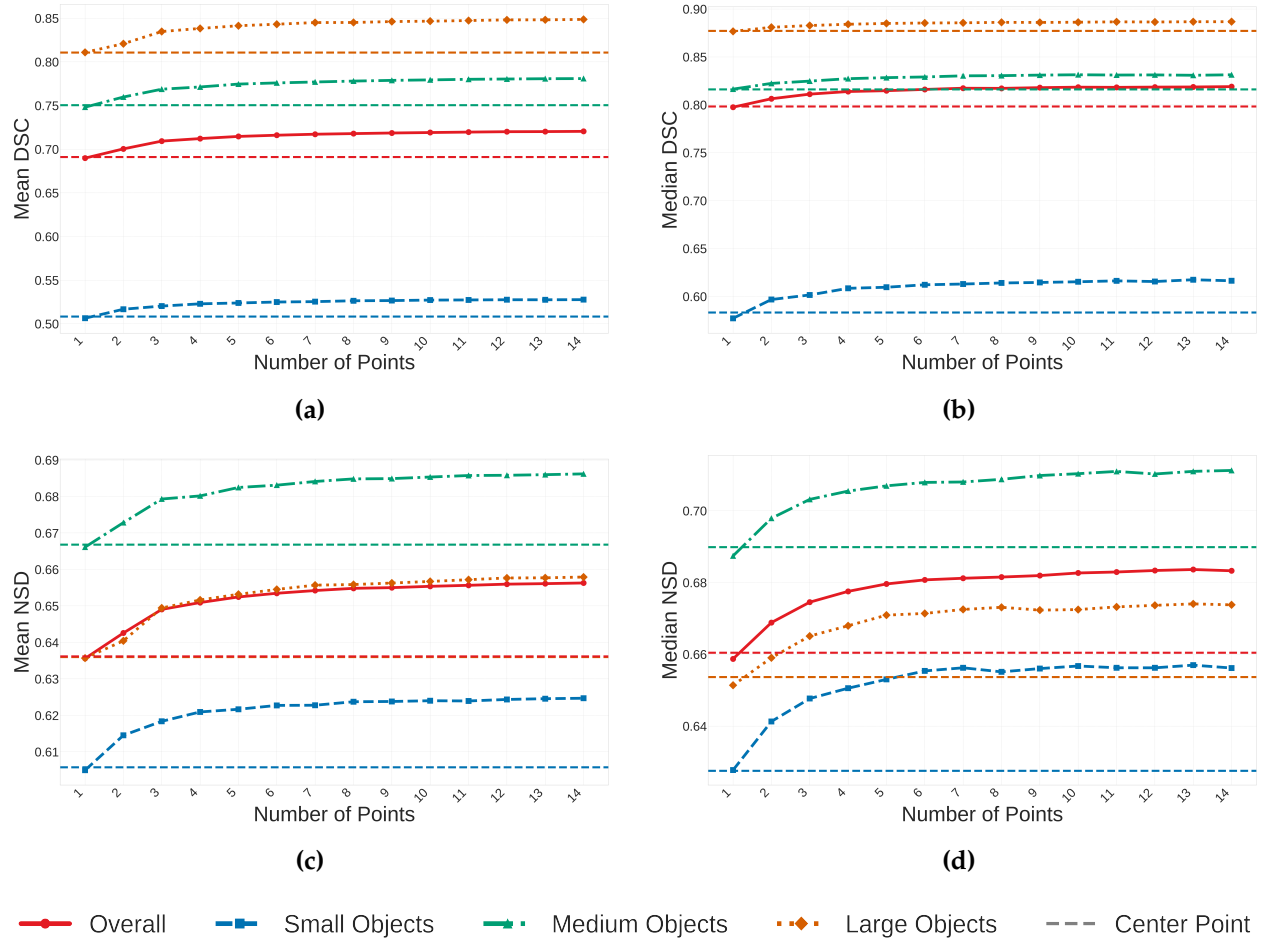


Figure S7: Mean Aggregation analysis for MSD_BrainTumor. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

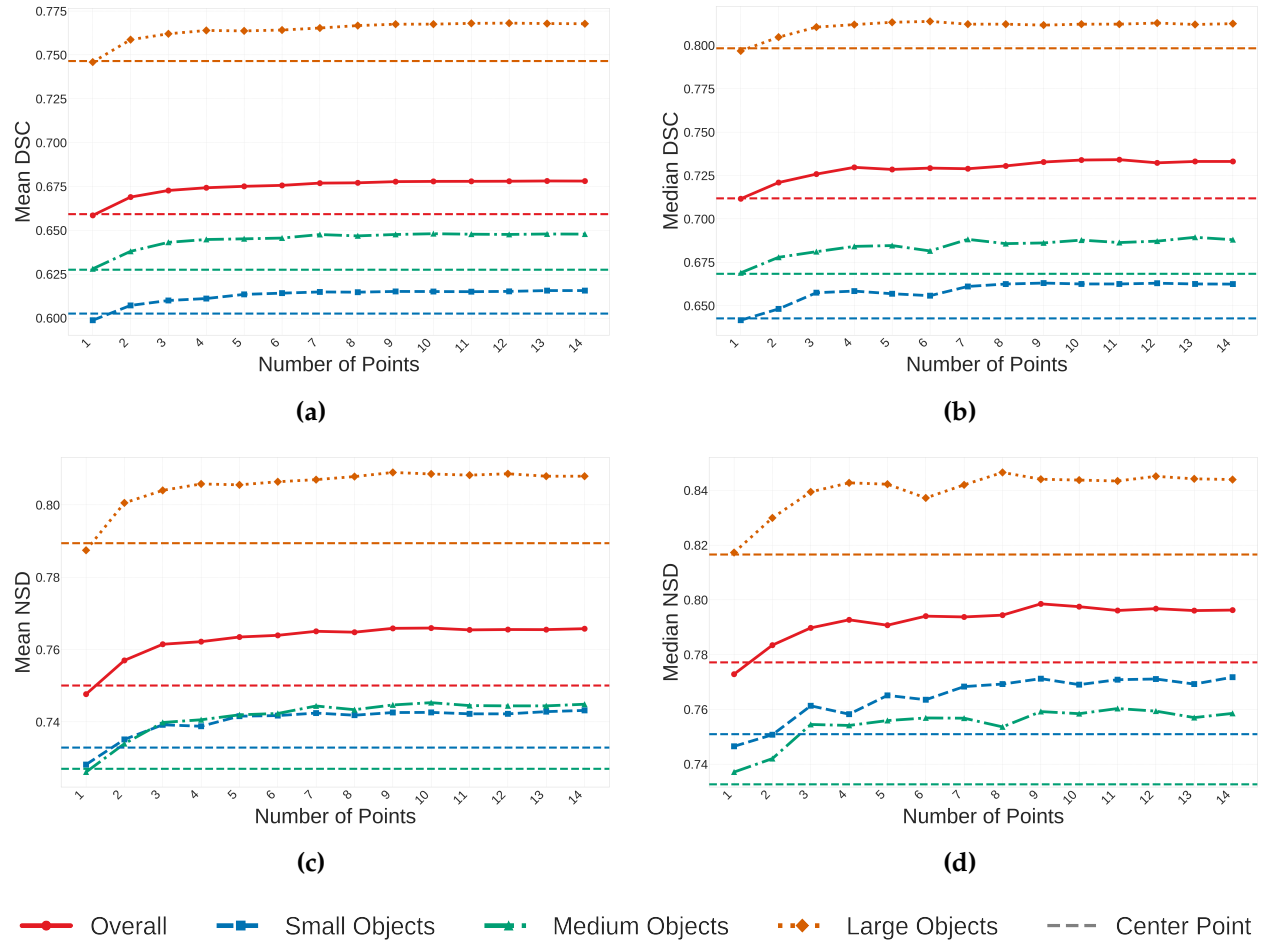


Figure S8: Mean Aggregation analysis for MsLesSeg_MSLesion. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

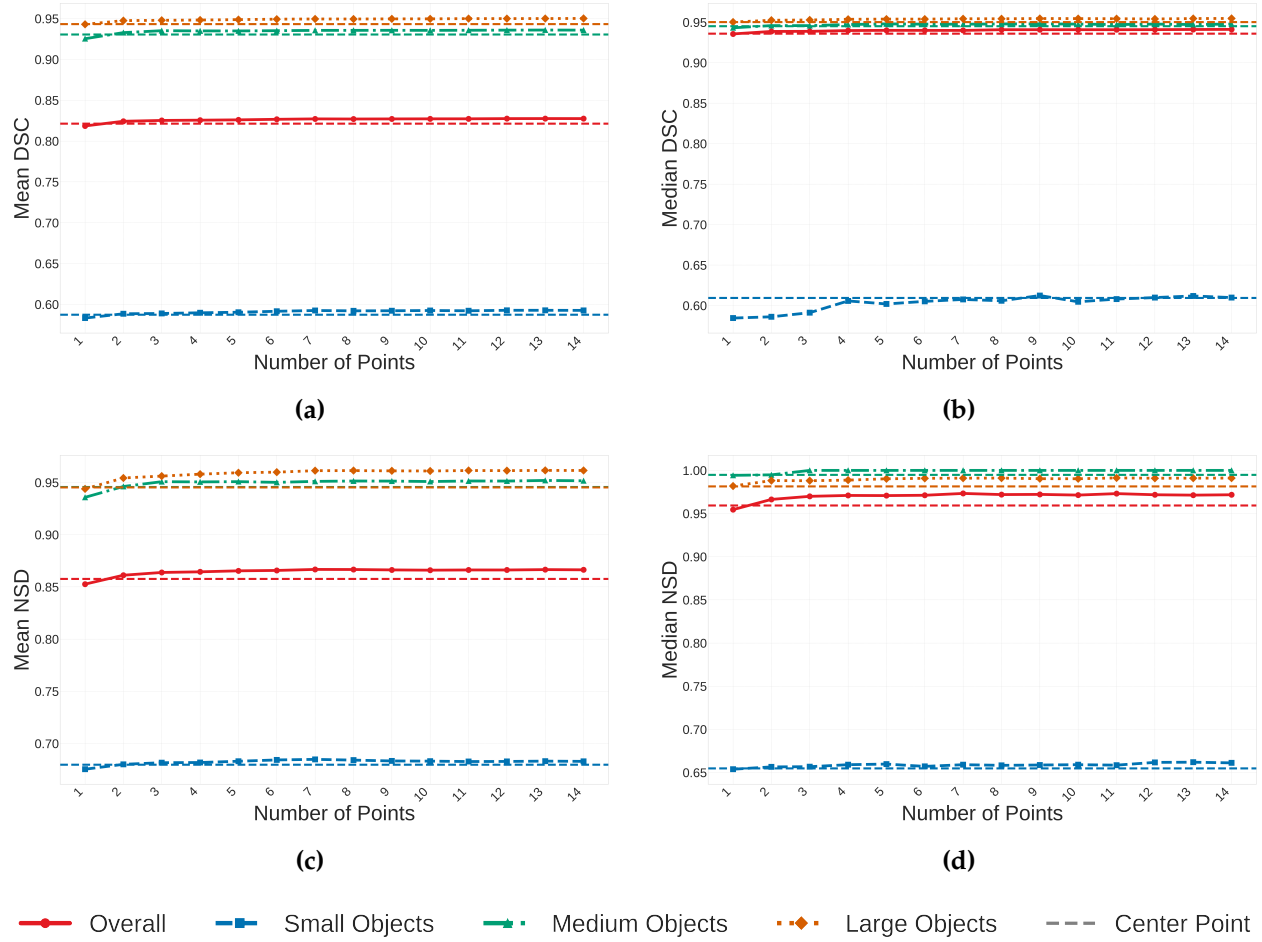


Figure S9: Mean Aggregation analysis for LASC_Heart. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

S3.2 Median Aggregation Method

This section presents the results obtained using the median aggregation strategy. Table S2 summarizes the performance metrics (DSC, NSD, and Inference Time) for the method across all eight datasets.

Table S2: Summary of Semi-Automatic GT-guided Prompt (Median Aggregation) results across all metrics.

Dataset	Mean DSC	Median DSC	Mean NSD	Median NSD	Mean Time	Median Time
MSD_Spleen	0.9212 (0.1063)	0.9513 (0.1063)	0.7798 (0.1564)	0.8206 (0.1564)	0.2899 (0.0853)	0.2933 (0.0853)
MSD_Liver	0.8535 (0.2584)	0.9590 (0.2584)	0.6952 (0.2274)	0.7476 (0.2274)	0.2481 (0.0699)	0.2509 (0.0699)
MSD_Pancreas	0.7590 (0.1847)	0.8218 (0.1847)	0.5345 (0.2138)	0.5453 (0.2138)	0.2332 (0.0700)	0.2392 (0.0700)
MSD_Colon	0.5610 (0.2755)	0.6105 (0.2755)	0.4255 (0.2220)	0.4195 (0.2220)	0.4453 (0.1049)	0.4636 (0.1049)
KiTS_Kidney	0.8205 (0.2134)	0.9120 (0.2134)	0.8009 (0.2209)	0.8858 (0.2209)	0.2580 (0.2212)	0.1095 (0.2212)
MSD_BrainTumor	0.7195 (0.2517)	0.8173 (0.2517)	0.6550 (0.2057)	0.6823 (0.2057)	0.1800 (0.0761)	0.1664 (0.0761)
MsLesSeg_MSLesion	0.6761 (0.2171)	0.7316 (0.2171)	0.7640 (0.1894)	0.7967 (0.1894)	0.3448 (0.1092)	0.3213 (0.1092)
LASC_Heart	0.8265 (0.2303)	0.9394 (0.2303)	0.8660 (0.1862)	0.9713 (0.1862)	0.2869 (0.0926)	0.2704 (0.0926)

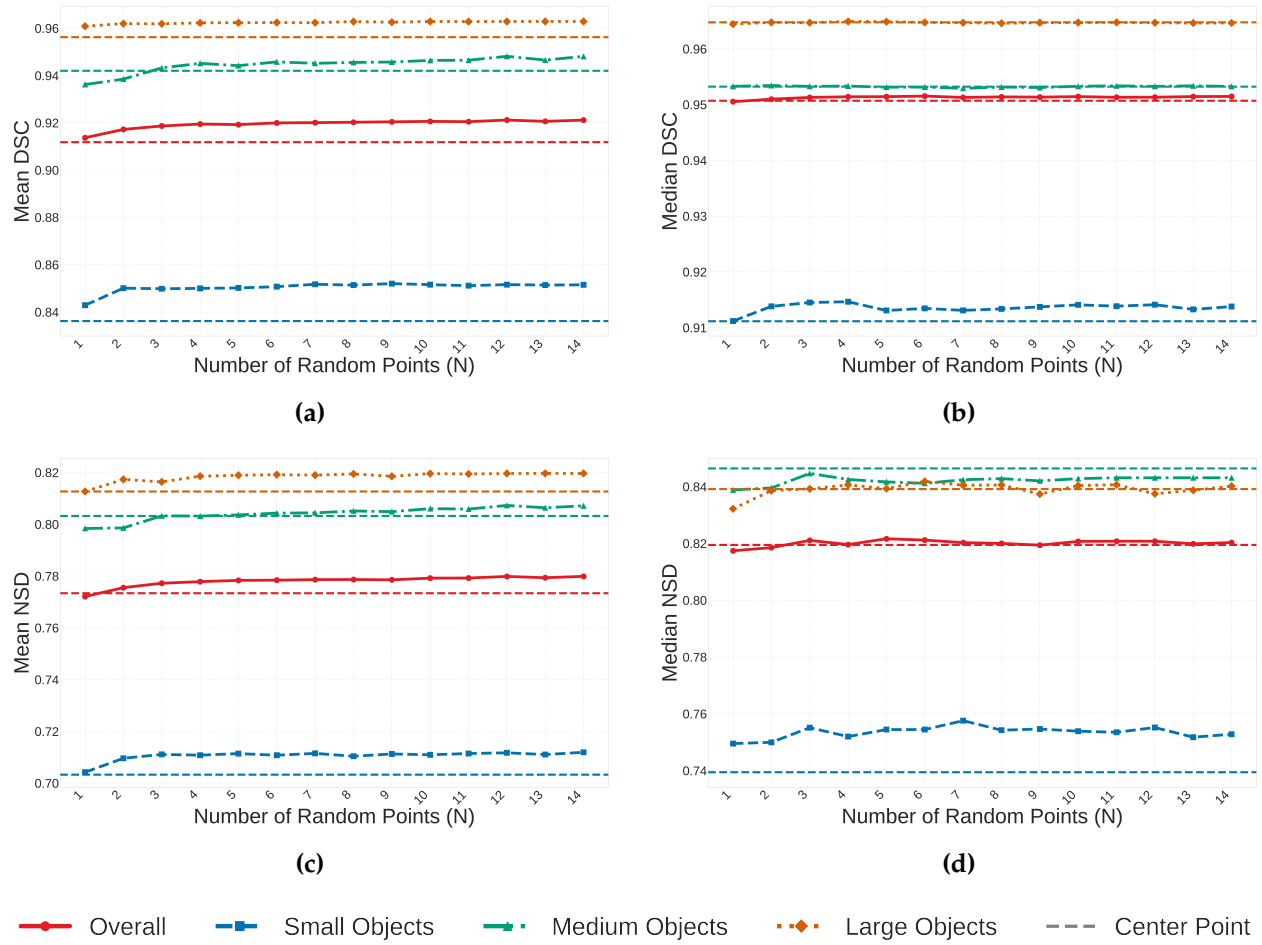


Figure S10: Median Aggregation analysis for MSD_Spleen. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

Figs. S10–S17 present the performance pattern using Median Aggregation strategy. These figures indicate the Mean and the Median Dice Similarity Coefficient (DSC) and the Normalized Surface Dice (NSD) of prompt points (N) of 0 to 14.

The results can be compared to the Mean Aggregation strategy in order to obtain similar overall trends, in which higher number of points tends to lead to a better performance. Median Aggregation is expected to exhibit a bit more robustness in the situations with extreme outliers, which is indicated by the consistency of measures when working with such datasets as MSD_Colon and MsLesSeg_MSLesion. The highest values of the DSC of Median Aggregation, however, at best are similar, and sometimes slightly less, than the highest values of Mean Aggregation. For instance, in the MSD_Pancreas dataset, Mean Aggregation reached a distinctly higher upper bound. This performance difference, together with the success of Mean Aggregation in smoothing variance, justified its use as the primary aggregation method in the main manuscript.

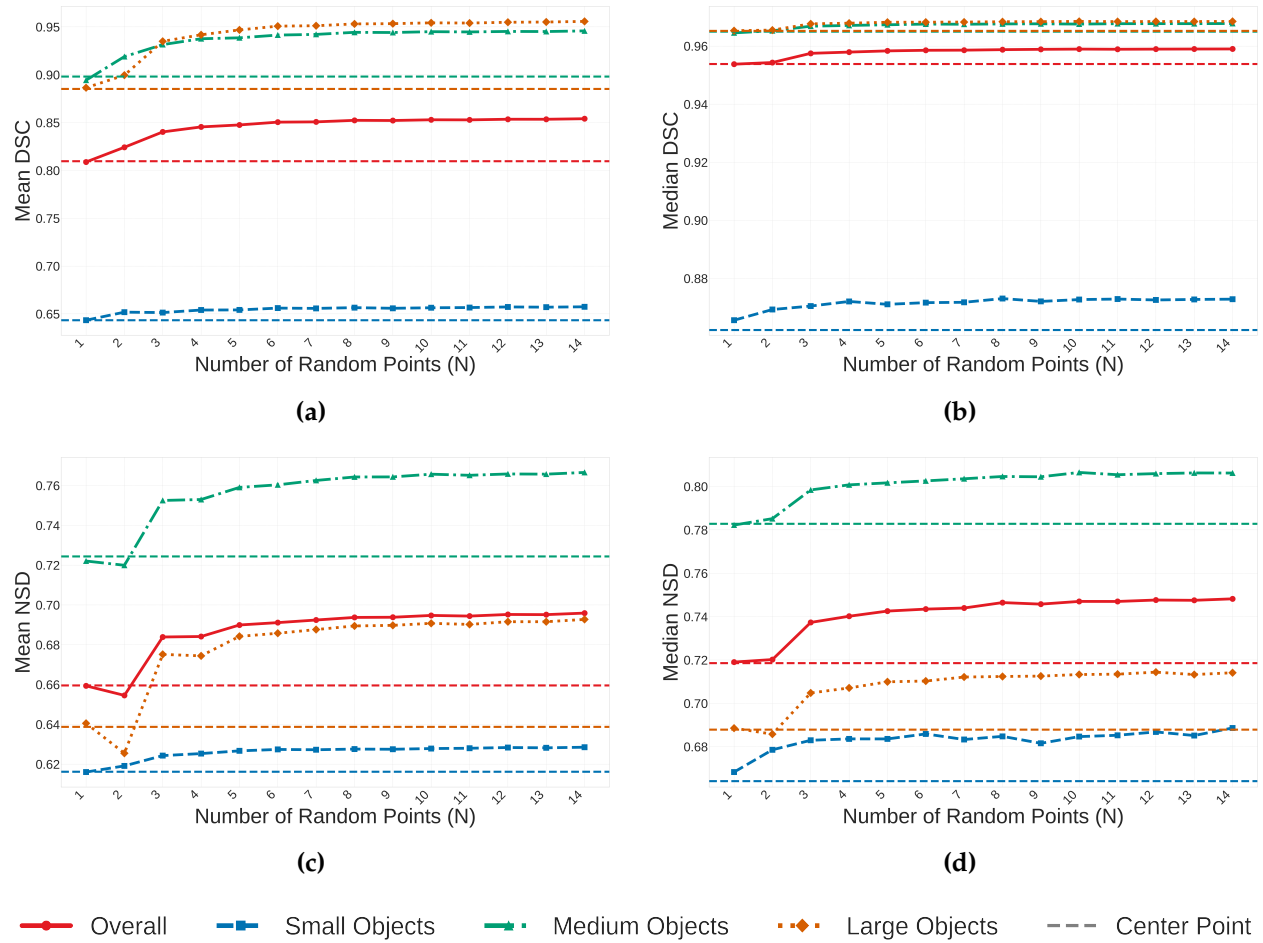


Figure S11: Median Aggregation analysis for MSD_Liver. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

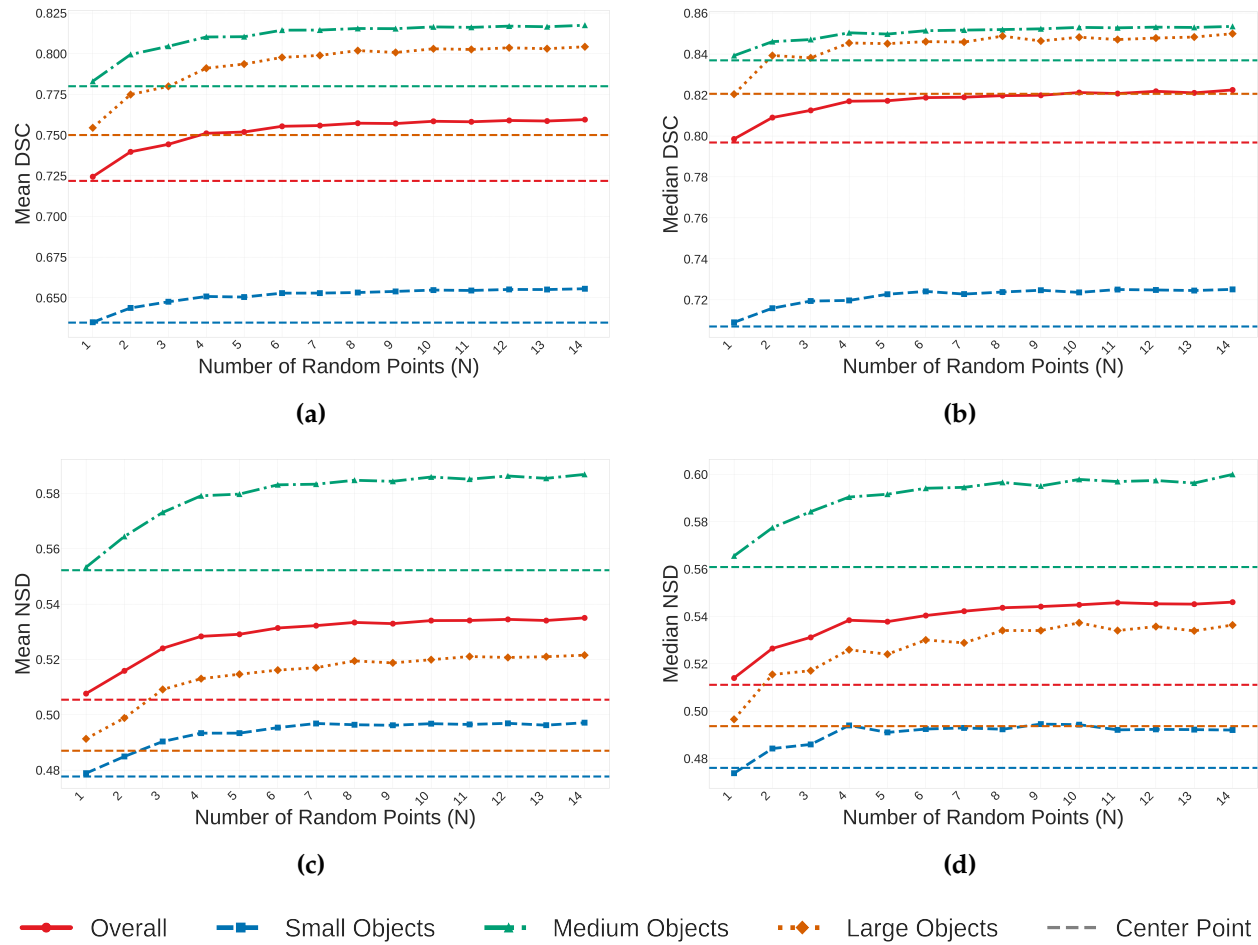


Figure S12: Median Aggregation analysis for MSD_Pancreas. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

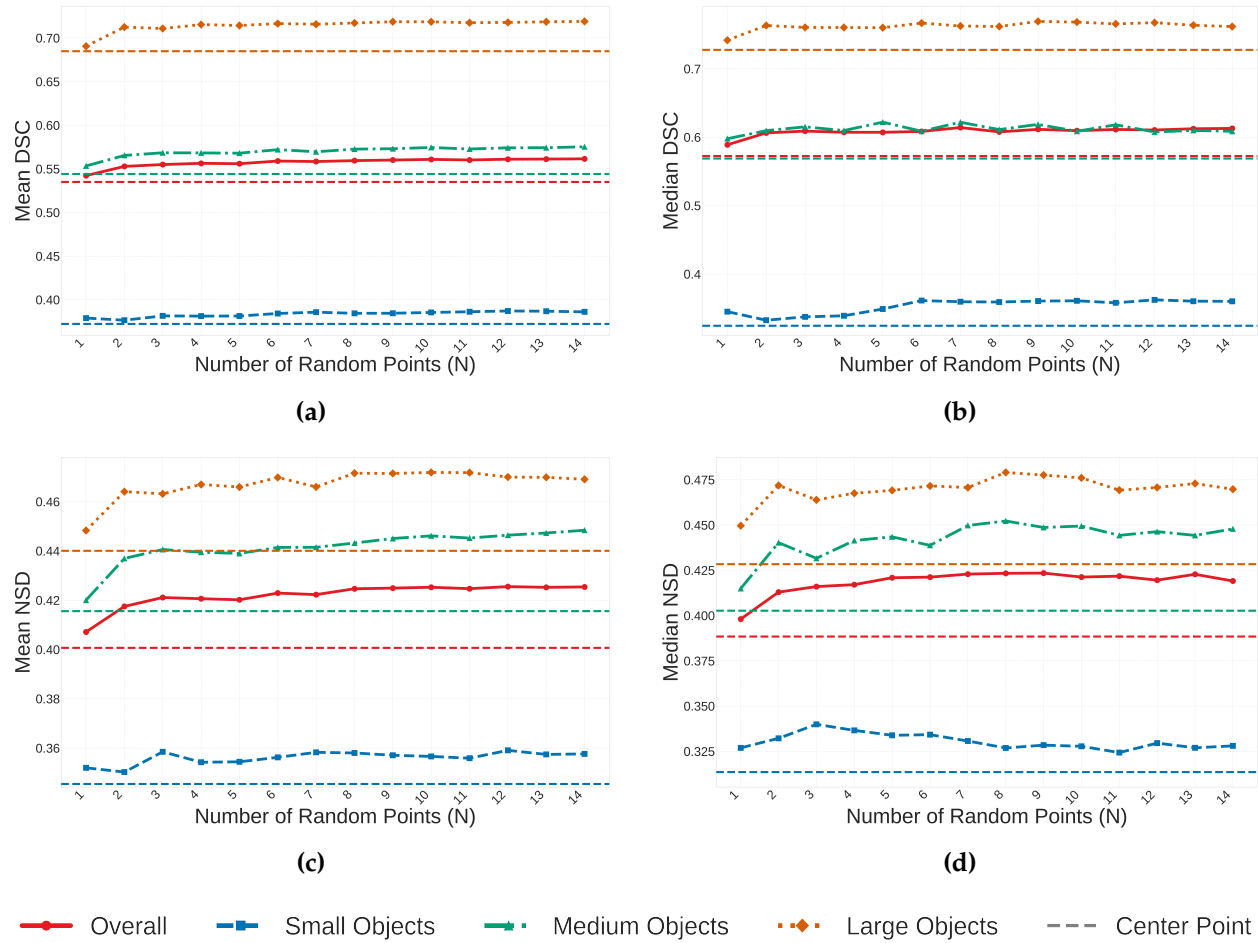


Figure S13: Median Aggregation analysis for MSD_Colon. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

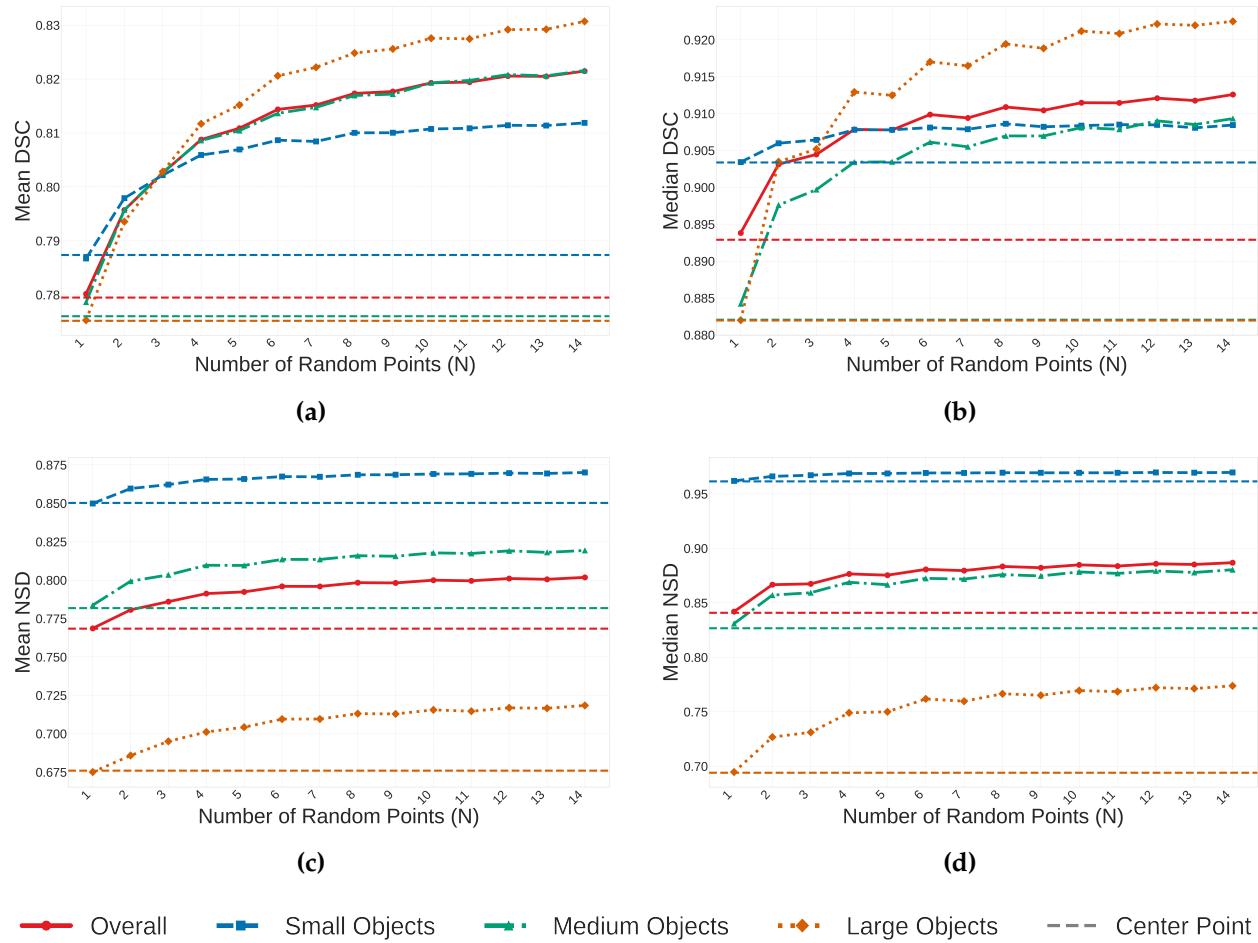


Figure S14: Median Aggregation analysis for KiTS_Kidney. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

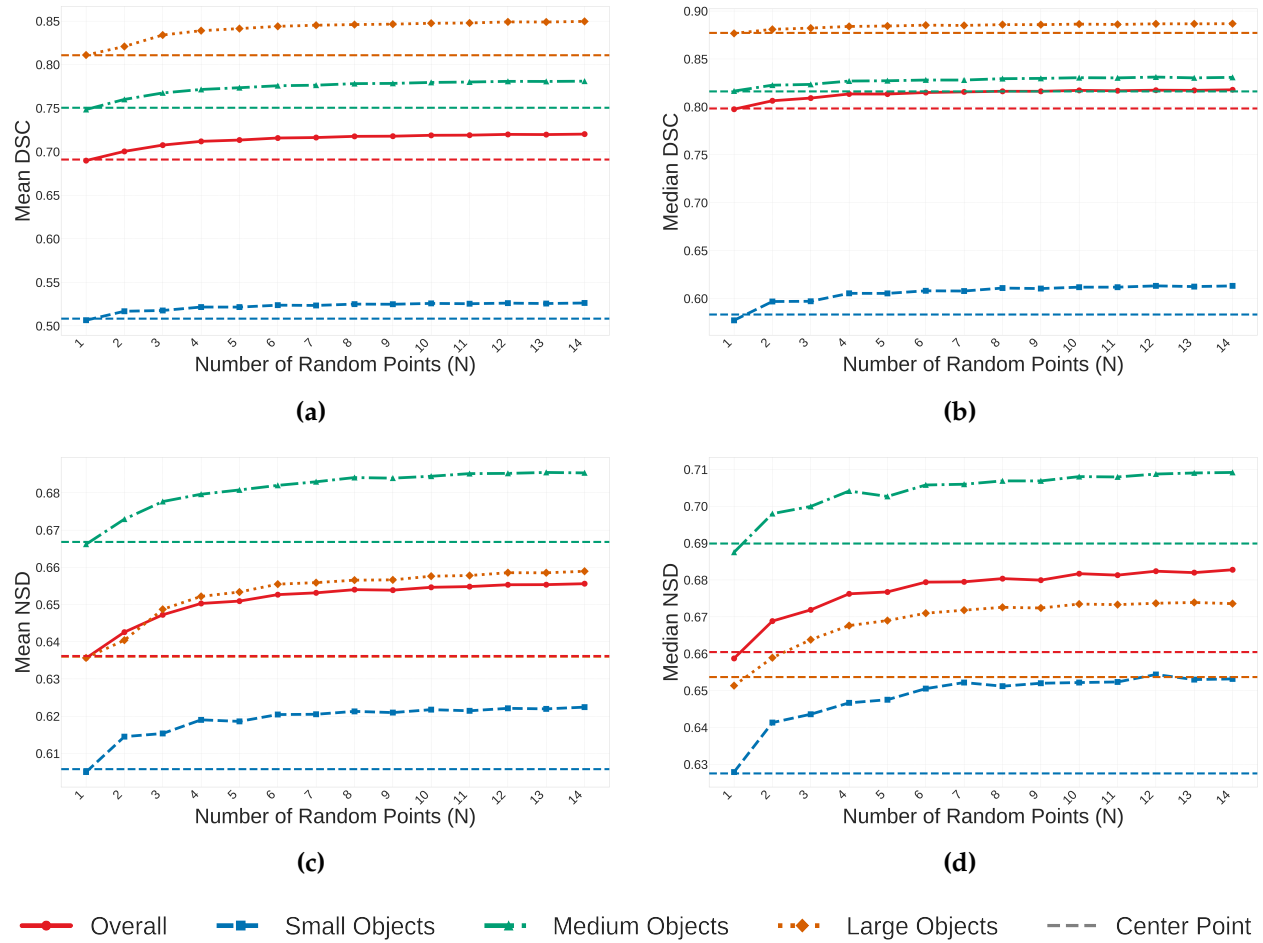


Figure S15: Median Aggregation analysis for MSD_BrainTumor. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

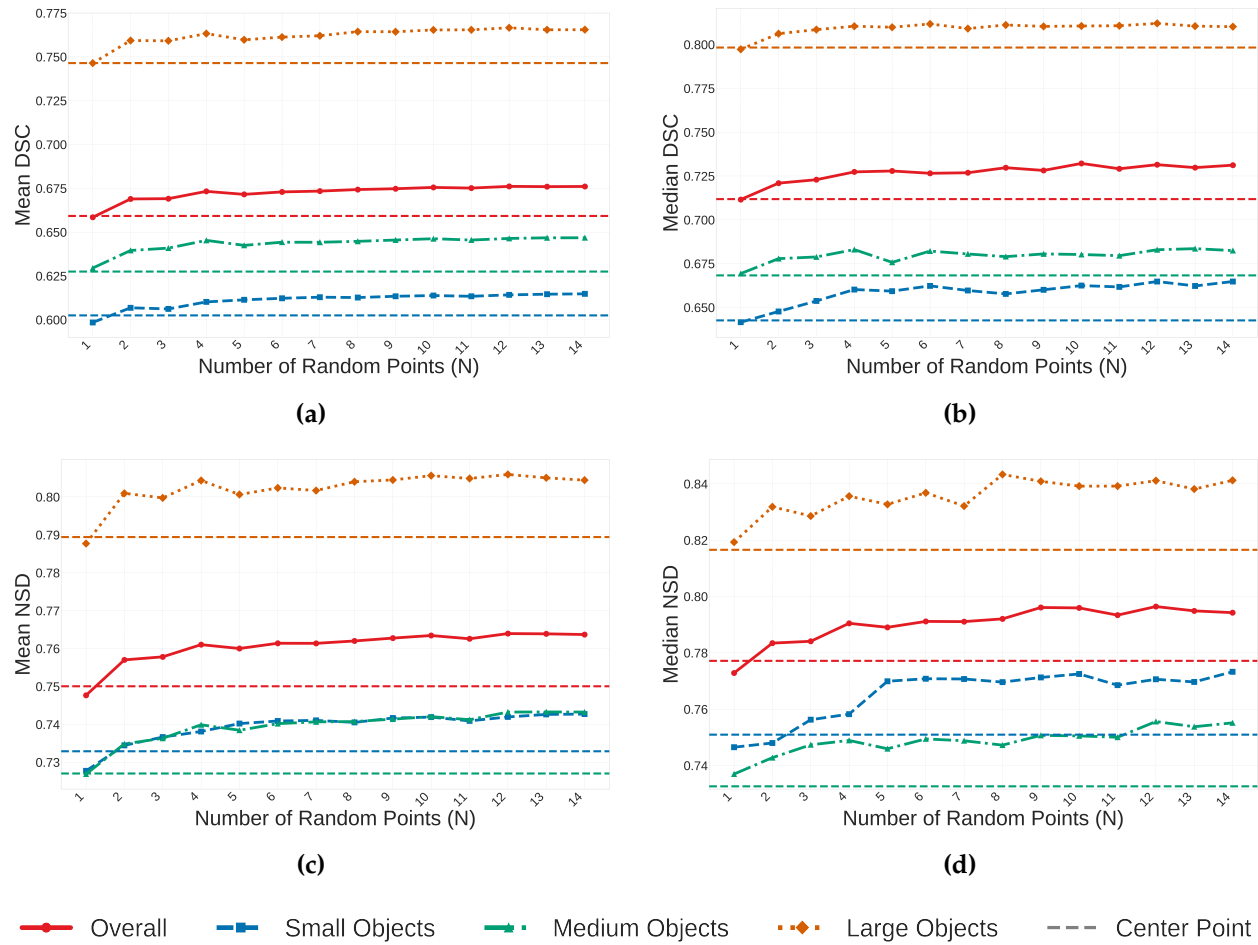


Figure S16: Median Aggregation analysis for MsLesSeg_MSLesion. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

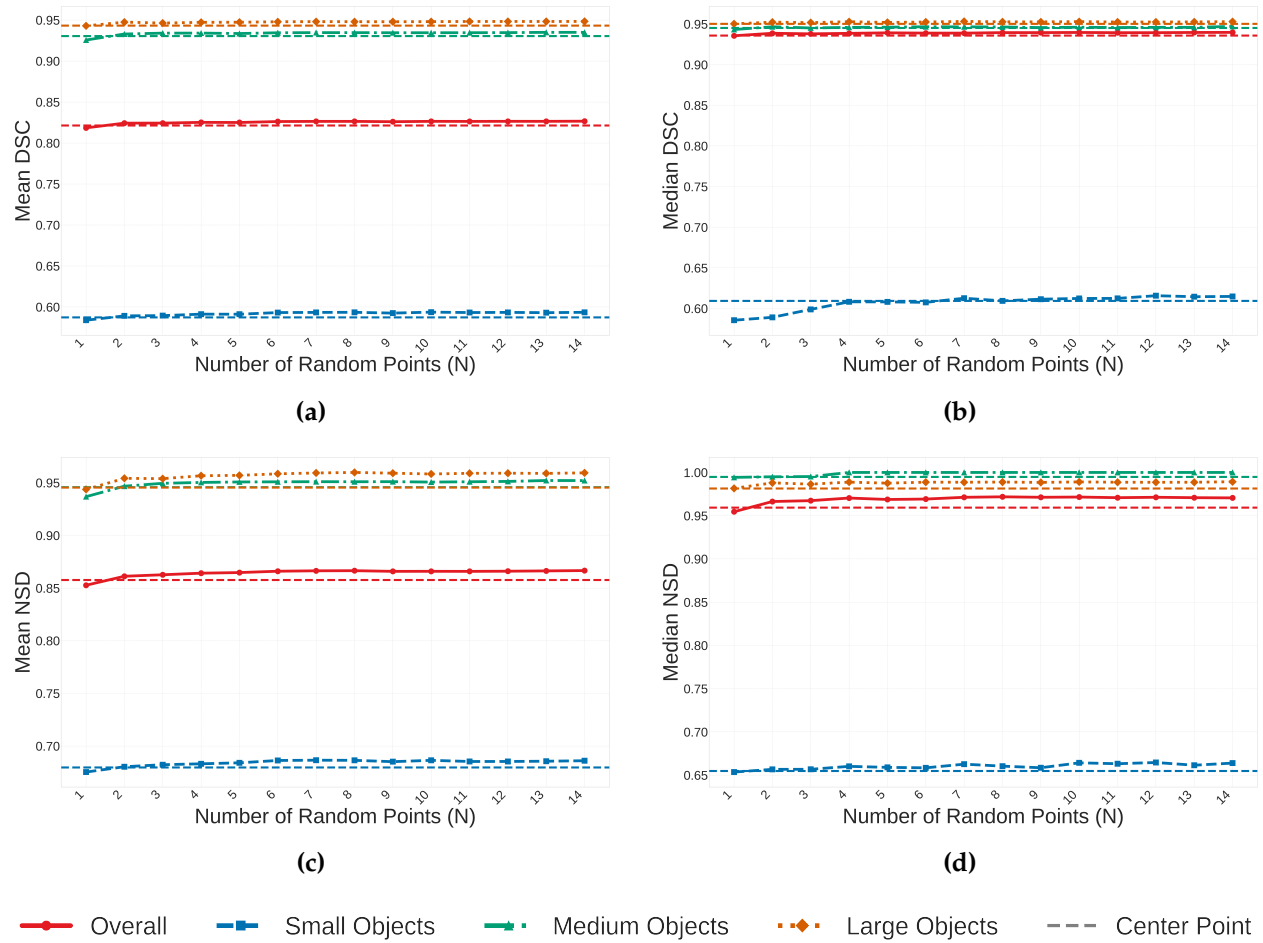


Figure S17: Median Aggregation analysis for LASC_Heart. (a) Mean DSC; (b) Median DSC; (c) Mean NSD; (d) Median NSD.

Table S3: Median DSC comparison of the proposed Multi-Point Aggregation strategies against SAM, Scribble Prompt, and the MedSAM (1-point) baseline.

Dataset	SAM	Scribble Prompt	MedSAM (1-point) baseline	Semi-Aut. GT-guided
MSD_Spleen	0.9079 (0.2798)	0.9054 (0.2866)	0.9507 (0.1299)	0.9519 (0.1067)
MSD_Liver	0.7645 (0.2791)	0.8415 (0.3516)	0.9539 (0.3044)	0.9590 (0.2576)
MSD_Pancreas	0.6081 (0.2763)	0.5155 (0.2383)	0.7969 (0.2147)	0.8237 (0.1853)
MSD_Colon	0.4273 (0.2962)	0.5677 (0.2385)	0.5725 (0.2765)	0.6207 (0.2771)
KiTS_Kidney	0.9384 (0.2656)	0.6965 (0.3087)	0.8929 (0.2411)	0.9132 (0.2155)
MSD_BrainTumor	0.1942 (0.2521)	0.3529 (0.2442)	0.7983 (0.2670)	0.8191 (0.2519)
MsLesSeg_MSLesion	0.0721 (0.3307)	0.6884 (0.2312)	0.7118 (0.2222)	0.7331 (0.2175)
LASC_Heart	0.7057 (0.2784)	0.7116 (0.2245)	0.9359 (0.2338)	0.9412 (0.2307)

Table S4: Median NSD comparison demonstrating boundary adherence of SAM, Scribble Prompt, and the MedSAM (1-point) baseline versus Multi-Point Aggregation.

Dataset	SAM	Scribble Prompt	MedSAM (1-point) baseline	Semi-Aut. GT-guided
MSD_Spleen	0.6117 (0.3056)	0.6155 (0.2347)	0.8195 (0.1617)	0.8221 (0.1562)
MSD_Liver	0.2885 (0.3138)	0.4242 (0.2515)	0.7186 (0.2524)	0.7468 (0.2289)
MSD_Pancreas	0.3570 (0.2403)	0.2966 (0.1868)	0.5111 (0.2180)	0.5442 (0.2181)
MSD_Colon	0.2214 (0.2540)	0.3612 (0.1945)	0.3884 (0.2178)	0.4245 (0.2250)
KiTS_Kidney	0.8875 (0.2907)	0.6253 (0.2853)	0.8409 (0.2301)	0.8904 (0.2256)
MSD_BrainTumor	0.0607 (0.2001)	0.4054 (0.2288)	0.6605 (0.2088)	0.6833 (0.2071)
MsLesSeg_MSLesion	0.1155 (0.3663)	0.7974 (0.1880)	0.7772 (0.1889)	0.7963 (0.1914)
LASC_Heart	0.5446 (0.3091)	0.5464 (0.2337)	0.9594 (0.1898)	0.9718 (0.1869)

S3.3 Quantitative analysis with Mean Aggregation function

Table S3 details the Median Dice Similarity Coefficient (DSC) comparisons across all datasets. The proposed Semi-Automatic GT-guided method consistently outperforms the standard SAM and Scribble Prompt approaches. Notably, in challenging tasks with irregular boundaries such as MSD_Colon and MsLesSeg_MSLesion, the proposed method achieves median DSCs of 0.6207 and 0.7331 respectively, significantly surpassing the Scribble Prompt scores. Furthermore, the method shows distinct improvements over the MedSAM (1-point) baseline, particularly in LASC_Heart (0.9412 vs 0.9359) and MSD_BrainTumor (0.8191 vs 0.7983). Table S4 presents the Normalized Surface Dice (NSD) metrics to assess boundary adherence. The multi-point aggregation strategy yields tighter boundary conformity than the baselines, evidenced by the high median NSD scores in MSD_Spleen (0.8221) and KiTS_Kidney (0.8904). Comparisons in the MSD_Liver dataset further validate this improvement, where the proposed method achieves a median NSD of 0.7468 compared to the single-point baseline of 0.7186.

S3.4 Qualitative analysis with Mean Aggregation function

Fig. S18 illustrates the qualitative segmentation performance across four representative datasets: MSD_Colon, MSD_BrainTumor, MsLesSeg_MSLesion, and LASC_Heart. The visual comparison highlights the limitations of the standard SAM and Scribble Prompt methods, which frequently exhibit over-segmentation or fail to capture complex, irregular boundaries. In contrast, the Semi-Automatic GT-guided method produces segmentation masks that align closely with the Ground Truth. For instance, in the MSD_BrainTumor samples (second row), the proposed method

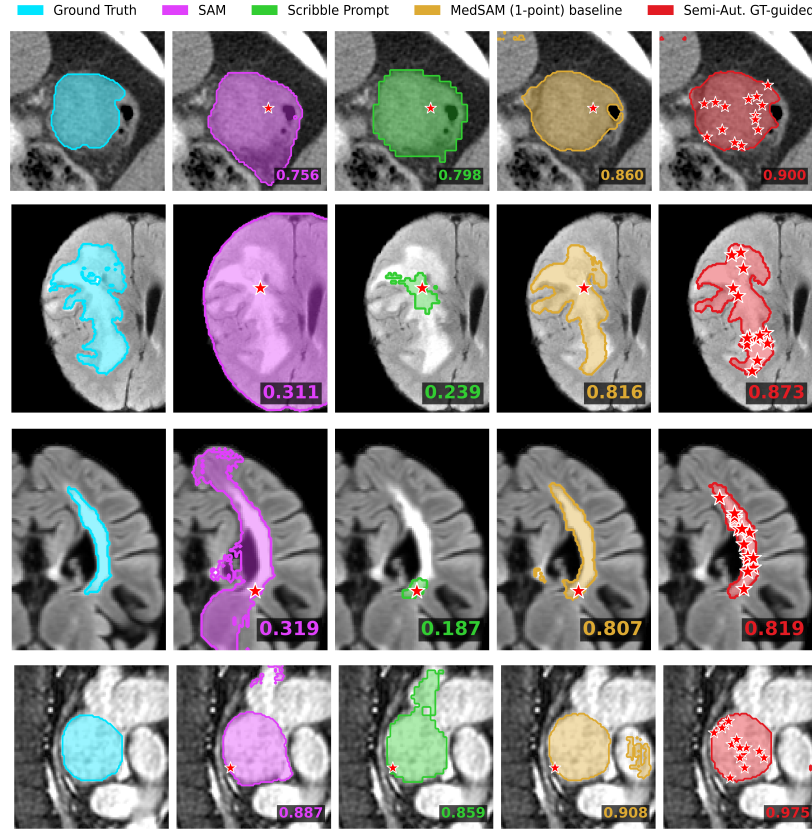


Figure S18: Qualitative results for Experiment 1. Rows correspond to (top to bottom): MSD_Colon, MSD_BrainTumor, MsLesSeg_MSLEsion, and LASC_Heart.

successfully delineates the complex tumor structure, whereas the Scribble Prompt and SAM baselines significantly under-segment or misinterpret the region of interest.

S4 Experiment 2: Automatic MedSAM-guided Prompt

This section presents the additional results for the fully automatic pipeline, categorizing the findings by the aggregation strategy used (Mean vs. Median).

S4.1 Mean Aggregation Method

Fig. S19 presents the Dice Similarity Coefficient (DSC) distributions for the Mean Aggregation strategy, as referenced in the main manuscript. In the MSD_Pancreas, KiTS_Kidney, and MsLesSeg_MSLEsion datasets, the median DSC of the Automatic MedSAM-guided method is higher than the median DSC of the MedSAM (1-point) baseline. In MSD_Pancreas, the interquartile range of the Automatic method is higher than the baseline range. In MSD_Spleen and MSD_Liver, the median DSC values for all methods are above 0.90.

S4.2 Median Aggregation Method

This section provides the results for the fully automatic pipeline using the Median Aggregation strategy, compared to the single-point baseline and the semi-automatic method.

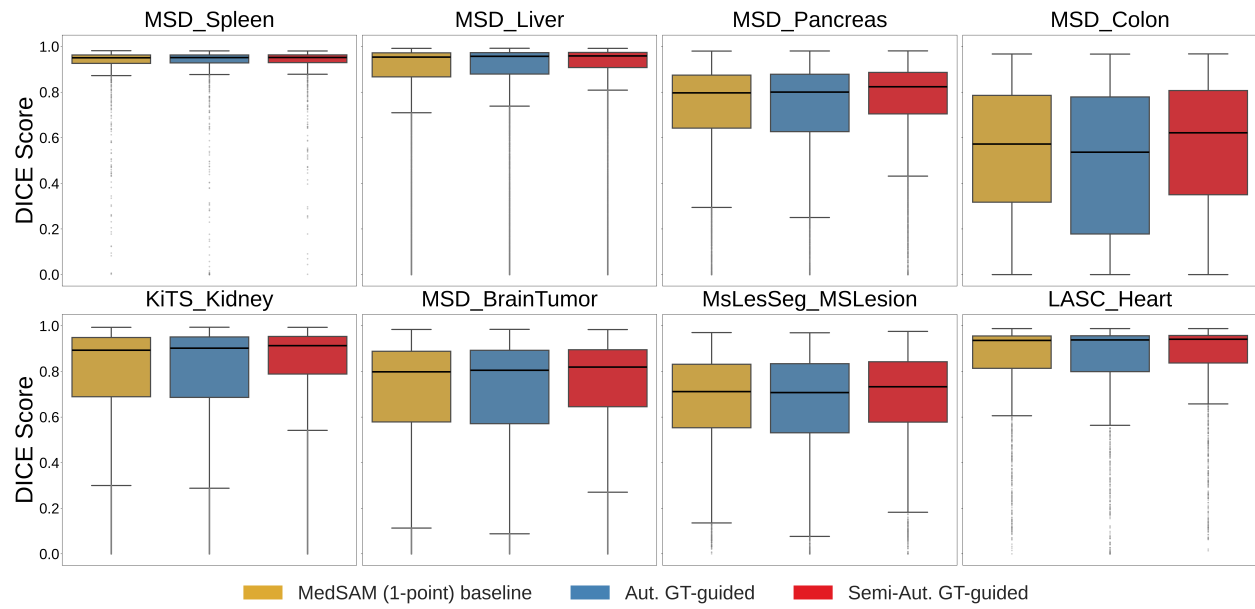


Figure S19: Dice Similarity Coefficient (DSC) distributions for Mean Aggregation function across all datasets.

DSC distributions for Median Aggregation (Fig. S20) show similar results to the Mean Aggregation findings in the main paper. The Automatic method achieves comparable median DSC to the MedSAM (1-point) baseline in MSD_Pancreas, KiTS_Kidney, and MsLesSeg_MSLesion. Distributions for MSD_Spleen, MSD_Liver, and LASC_Heart are tightly clustered with high median values.

NSD distributions (Fig. S21) show the same trend. The Automatic method outperforms the Baseline in median NSD for MSD_Pancreas and KiTS_Kidney. MSD_Colon shows high variance across all methods.

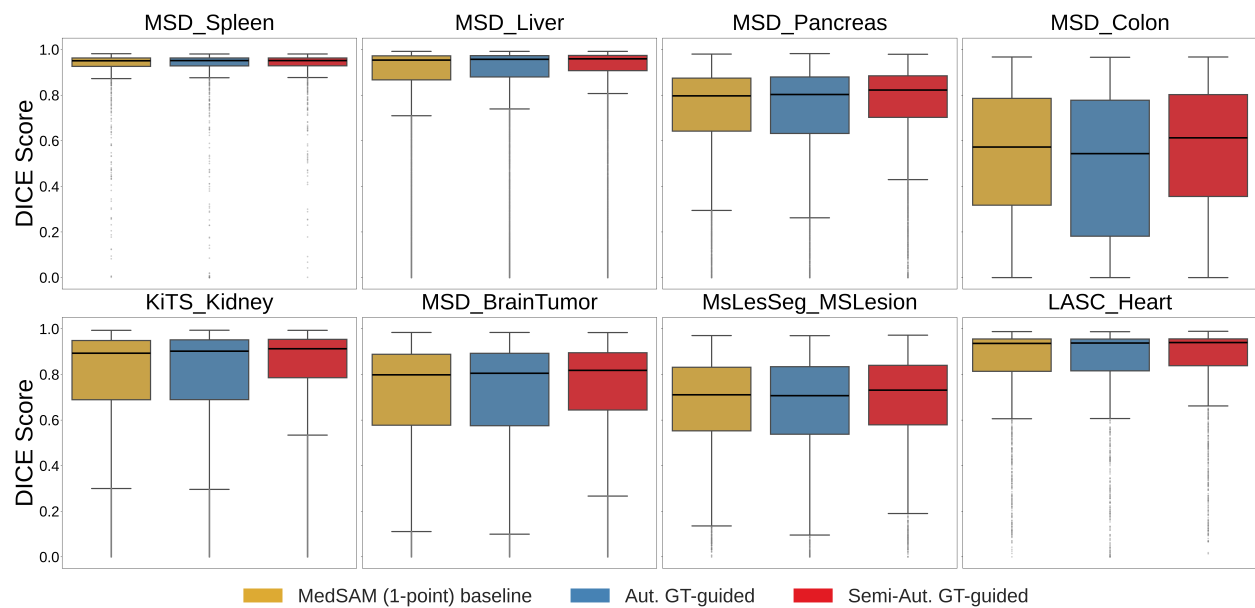


Figure S20: Dice Similarity Coefficient (DSC) distributions for Median aggregation function across all datasets.

Table S5: Mean Inference Time (s) comparison. MedSAM (1-point) represents the fastest baseline, compared against SAM, Scribble Prompt, and the proposed Multi-Point Aggregation.

Dataset	SAM	Scribble Prompt	MedSAM (1-point) baseline	Semi-Aut. GT-guided
MSD_Spleen	0.1352 (0.0259)	0.5368 (0.2492)	0.1342 (0.0258)	0.4618 (0.0968)
MSD_Liver	0.0328 (0.0007)	0.9377 (0.0742)	0.0327 (0.0008)	0.0967 (0.0043)
MSD_Pancreas	0.0635 (0.0196)	0.3487 (0.1534)	0.0632 (0.0195)	0.2436 (0.0881)
MSD_Colon	0.1374 (0.0420)	1.2092 (0.4965)	0.1365 (0.0420)	0.4166 (0.1442)
KiTS_Kidney	0.0721 (0.0226)	0.8904 (0.2721)	0.0716 (0.0226)	0.1618 (0.0811)
MSD_BrainTumor	0.0327 (0.0009)	0.8223 (0.3168)	0.0325 (0.0011)	0.0910 (0.0066)
MsLesSeg_MSLesion	0.0948 (0.0173)	0.5818 (0.1019)	0.0944 (0.0202)	0.2817 (0.0679)
LASC_Heart	0.0744 (0.0212)	1.2789 (0.0213)	0.0745 (0.0217)	0.3213 (0.0968)

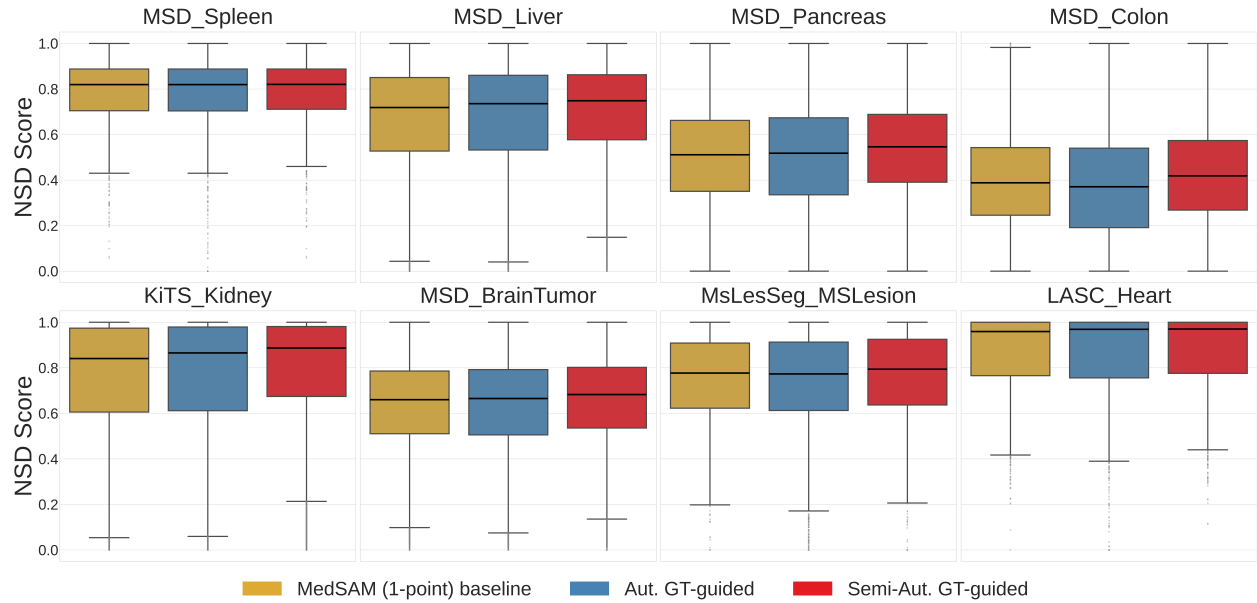


Figure S21: Normalized Surface Dice (NSD) distributions for Median aggregation function across all datasets.

S5 Inference time

Tables S5 and S6 list the inference times. The single-point MedSAM baseline ranged from 0.03s to 0.14s. The Semi-Automatic GT-guided Prompt ($N = 14$) ranged from 0.09s to 0.46s. Scribble Prompt inference times were 1.2092s for MSD_Colon and 0.9377s for MSD_Liver.

In Table S5 the median inference time is summarized and shows efficiency trade-offs between the methods. Although the Semi-Automatic GT-guided process slightly increases the amount of time that a particular computation takes than the single-point MedSAM baseline (e.g., 0.4697s vs 0.1337s to complete the MSD_Spleen step), it is still much faster than the Scribble Prompt method, which takes 0.5614s to do the same operation. This indicates that the proposed strategy of aggregation is advantageous in terms of performance because it gives a better segmentation and incurring a small cost in the inference latency.

Table S6: Median Inference Time (s) comparison across SAM, Scribble Prompt, MedSAM (1-point), and Multi-Point Aggregation.

Dataset	SAM	Scribble Prompt	MedSAM (1-point) baseline	Semi-Aut. GT-guided
MSD_Spleen	0.1353 (0.0259)	0.5614 (0.2492)	0.1337 (0.0258)	0.4697 (0.0968)
MSD_Liver	0.0327 (0.0007)	0.9220 (0.0742)	0.0326 (0.0008)	0.0955 (0.0043)
MSD_Pancreas	0.0576 (0.0196)	0.2683 (0.1534)	0.0568 (0.0195)	0.2377 (0.0881)
MSD_Colon	0.1311 (0.0420)	1.5973 (0.4965)	0.1301 (0.0420)	0.3946 (0.1442)
KiTS_Kidney	0.0680 (0.0226)	0.9248 (0.2721)	0.0681 (0.0226)	0.1479 (0.0811)
MSD_BrainTumor	0.0325 (0.0009)	0.5795 (0.3168)	0.0323 (0.0011)	0.0888 (0.0066)
MsLesSeg_MSLesion	0.1044 (0.0173)	0.5439 (0.1019)	0.1039 (0.0202)	0.2878 (0.0679)
LASC_Heart	0.0804 (0.0212)	1.2781 (0.0213)	0.0810 (0.0217)	0.3100 (0.0968)

References

1. Ma J, Zhang Y, Gu S, Ge C, Ma S, Young A, et al. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the FLARE22 challenge. *Lancet Digit Health*. 2024;6(2):e113-24.
2. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation decathlon. *Nat Commun*. 2022;13(1):4128.
3. Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, Li F, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. *Med Image Anal*. 2021;67:101821.
4. Guarnera F, Rondinella A, Crispino E, Russo G, Di Lorenzo C, Maimone D, et al. MSLesSeg: baseline and benchmarking of a new Multiple Sclerosis Lesion Segmentation dataset. *Sci Data*. 2025;12(1):920.
5. Simpson AL, Antonelli M, Bakas S, Bilello M, Farahani K, Van Ginneken B, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv:1902.09063*. 2019.