iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** II **Month of publication:** February 2026

DOI: https://doi.org/10.22214/ijraset.2026.77569

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue II Feb 2026- Available at www.ijraset.com

Zero-Shot 3D Brain Tumor Segmentation: Evaluating SAM2 across Multimodal Brain MRI

Tommy Nicholas[1], Louis Bratawidjaja[2]
School of Artificial Intelligence, Nanjing University of Information Science and Technology

Abstract: ***Brain tumor segmentation from MRI is clinically important yet challenging to deploy reliably, since tumor appearance varies widely across patients and boundary visibility depends strongly on the MRI modality. In this work, we evaluate Segment Anything Model 2 (SAM2) as a training-free, prompt-driven approach for volumetric tumor delineation in a zero-shot setting on the BraTS2024 Glioma Post-treatment dataset. To reflect a practical inference workflow, we apply minimal preprocessing and generate an initial 2D mask on a selected slice before predicting the full volume using bidirectional slice propagation with state reset to limit drift. We benchmark eight prompting strategies spanning point-only prompts, bounding box prompts, and box-plus-point combinations, and compare performance across four modalities (T1n, T1ce, T2w, and FLAIR) using IoU and Dice on a binary tumor mask derived from the original multi-class annotations. The results show that prompt design substantially influences both accuracy and stability, with box-guided prompting consistently outperforming point-only interaction and additional positive points further improving robustness. We also observe a clear modality effect, where FLAIR and T2w provide more reliable delineation cues than T1-based modalities under the same prompting and propagation protocol. These findings clarify when SAM2 is dependable for zero-shot volumetric tumor segmentation and provide practical guidance on prompt selection and modality choice for interactive clinical use.***

Keywords: SAM2, 3D segmentation, Zero-shot segmentation, Medical Image Analysis, Brain tumor segmentation

I. INTRODUCTION

Brain tumor segmentation plays a central role in diagnosis, treatment planning, and longitudinal monitoring, yet it remains one of the most demanding tasks in medical image analysis because the target is not a single uniform object but a set of heterogeneous subregions whose appearance shifts across patients, scanners, and acquisition protocols. Even within the same clinical category, tumors may present as compact masses, infiltrative patterns, or scattered enhancing foci, and these variations translate into inconsistent boundary contrast and ambiguous transitions between tumor and surrounding tissue. Deep learning has greatly advanced automated segmentation, but most supervised pipelines still depend on large curated annotations, extensive compute, and careful tuning of preprocessing and training recipes, and their reliability can fluctuate when the input modality changes or when the visual cues that a model has learned to rely on become muted or distorted. As a result, it is difficult to treat “brain tumor segmentation” as a single stable target across MRI, since what looks separable in one modality may be nearly indistinguishable in another, creating a practical gap between promising benchmark performance and a workflow that generalizes smoothly across routine clinical variability [1]. This variability is tightly linked to the fact that MRI is inherently multi-modality, where native T1-weighted imaging (T1n), contrast-enhanced T1-weighted imaging (T1ce), T2-weighted imaging (T2w), and T2-weighted fluid-attenuated inversion recovery (FLAIR) each emphasize different tissue properties and therefore surface different tumor cues [2]. T1ce often sharpens enhancing components, T1n provides anatomical structure without contrast, while T2w and FLAIR are particularly informative for edema and broader abnormal tissue extent. To further clarify the MRI modality appearances, we present the illustration in Fig. 1. This modality-dependent visibility naturally motivates segmentation approaches that do not hard-code a single set of learned cues into a fixed model, but instead can flexibly leverage whichever signal is most informative in the current input. Promptable foundation models align with this need by allowing the user to specify intent through lightweight interactions, and Segment Anything Model 2 (SAM2) operationalizes this idea by producing masks from simple prompts without task-specific retraining, making it a plausible candidate for training-free delineation across heterogeneous MRI modalities [3].

Unlike conventional models that require task-specific training with extensive labeled data, SAM2 enables prompt-driven segmentation that can be applied without retraining. However, its effectiveness in volumetric brain tumor MRI has not been systematically characterized across MRI modalities, prompt modes, and slice-wise propagation behavior.

In this study, we therefore evaluate SAM2 in a zero-shot setting on 3D brain tumor MRI and investigate:

1) Modality-dependent performance,
2) The relative effectiveness of prompt modes and combinations, and
3) Propagation consistency across slices in a bi-directional workflow.

We further provide a modality-stratified evaluation to investigate the performance in statistical way, a controlled prompt-strategy comparison, and a propagation stability analysis with representative cases to delineate the strengths and limitations of SAM2 for 3D tumor segmentation.

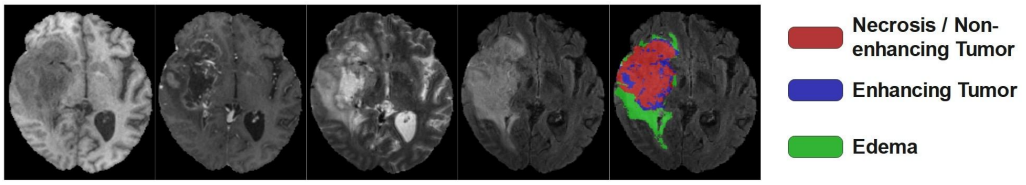


Fig. 1 The comparison of different brain MRI modalities. Each modality captures different tumor characteristics. The GT mask shows the different tumor sub-type generally used in clinical setting

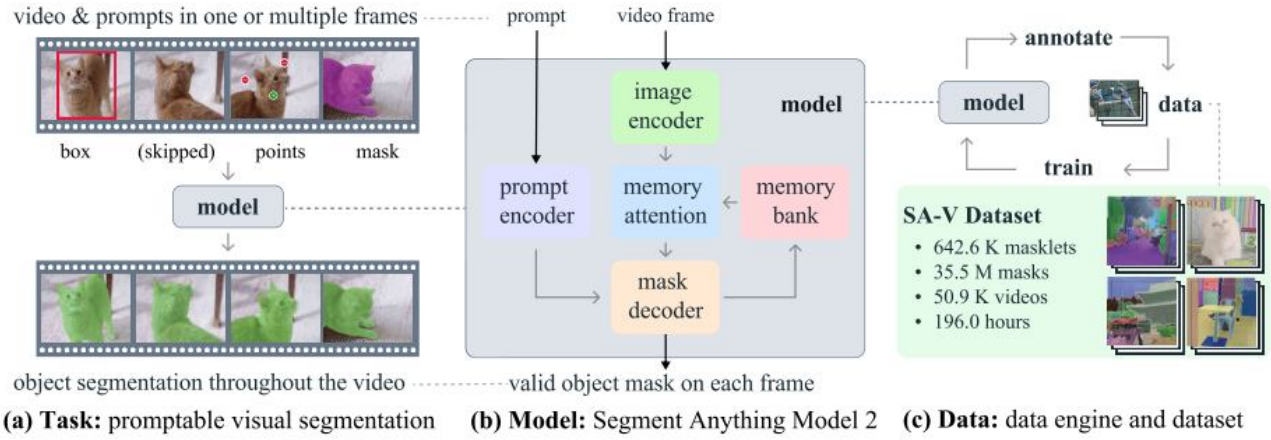


Fig. 2 SAM2 original process. Image is retrieved from the original paper [3]

II. RELATED WORKS

For years, researchers have relied on deep learning to advance medical image segmentation, motivated by its ability to learn rich appearance cues and delineate target boundaries with minimal hand-crafted rules. Among the many architectures proposed, U-Net remains a particularly durable backbone due to its encoder decoder design and skip connections that preserve spatial detail while aggregating contextual information [4–6]. In brain tumor segmentation, this structure has proven effective for recovering irregular boundaries and separating tumor regions from surrounding tissue. Building on this foundation, nnU-Net provides a strong and widely used starting point by standardizing a robust training pipeline around fully convolutional network [7]. However, purely convolutional designs can struggle to capture long-range dependencies across the full field of view, especially when tumor appearance is diffuse or when global anatomical context is needed to disambiguate boundaries [8]. This limitation has encouraged transformer-based segmentation models that explicitly model global interactions [9], as well as multi-modal fusion approaches that combine complementary MRI modalities to improve robustness under heterogeneous tumor presentations [10–12]. While these designs often improve performance, they typically require substantial computational resources and careful tuning to realize their gains, which can complicate practical deployment.

The Segment Anything Model (SAM) offers a different perspective on segmentation by shifting emphasis from task-specific training toward prompt-driven inference [13]. As a foundation model pretrained on large-scale data, SAM can generate masks conditioned on simple user prompts such as points or boxes, enabling flexible delineation without fine-tuning in many settings. Although SAM was originally developed for natural images, subsequent studies have explored its applicability to medical imaging, reporting evaluations across diverse datasets and anatomical targets, including brain tumors [14–17]. These works highlight the appeal of interactive segmentation for reducing annotation burden and enabling rapid, user-guided refinement.

At the same time, standard SAM operates in a 2D setting, which is a notable constraint for medical imaging workflows where volumetric consistency across slices is critical and where tumor morphology often extends across multiple planes. To address this gap, several extensions have been proposed for 3D usage, such as SAM3D and SAM-Med3D, which emphasize improved prompt encoding and adaptation strategies to better support volumetric inputs [18,19].

Following the growing demand for promptable segmentation in volumetric settings, Meta introduced SAM2, which extends SAM with a video-oriented design that incorporates temporal style memory and propagation mechanisms [20]. This formulation is especially relevant to 3D medical imaging because a volume can be treated as an ordered stack of slices, analogous to frames in a video, allowing prompts provided on selected slices to be propagated through neighboring slices. Early studies have reported initial evidence of SAM2's zero-shot capability in medical contexts [20], yet its performance has not been systematically examined across different brain MRI modalities, where boundary visibility and tumor cues vary substantially. Moreover, SAM2 is fundamentally prompt-conditioned, so the choice of prompt mode and prompt combination can strongly influence both the quality of the initial delineation and the stability of slice-to-slice propagation. Prior work has explored point-based and box-based prompting strategies in general settings [21], but a controlled, head-to-head comparison specifically for 3D brain tumor segmentation remains limited. Hence, motivated by these gaps, our study evaluates SAM2 across brain MRI modalities and investigates which prompting strategies yield the most reliable volumetric delineation and propagation behavior in a zero-shot setting.

III.METHODOLOGY

In this section, we analyze the zero-shot segmentation process. The general process is divided into four major parts as visualized in Fig. 3. We start the process with selecting the frame, followed by prompt annotation. Then, SAM2 model uses this information to generate the predicted segmentation mask. Finally, we calculate the IoU and Dice to get the statistics result.

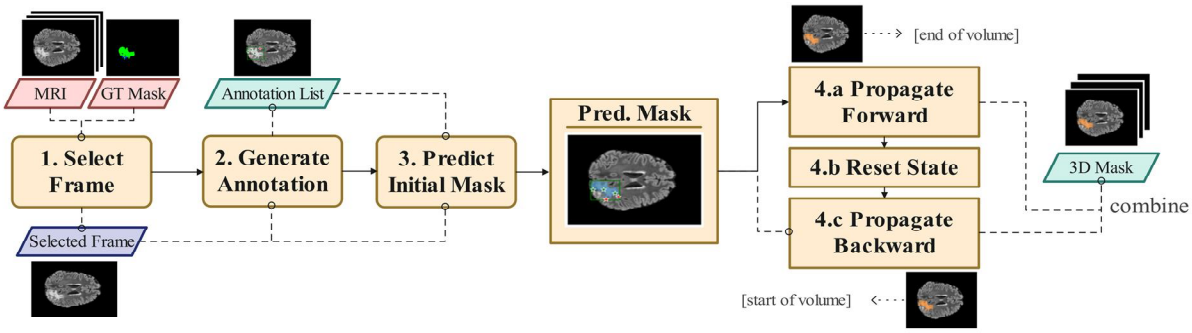


Fig. 3 The general process of the SAM2 evaluation strategy

A. *Select Frame*

The first step is to choose an initial slice from the 3D volume that serves as the starting point for prompting and subsequent propagation. This initialization is non-trivial because tumor burden can be unevenly distributed along the axial direction, and the most informative slice may differ substantially across patients and tumor phenotypes. One method is to directly take the middle slice of the brain MRI sample; however, it will make the tumor not properly highlighted across different samples. To reduce this variability and ensure a consistent selection protocol across the cohort, we determine the starting slice using the ground truth (GT) mask by identifying the slice that contains the largest foreground area of a prioritized tumor label. Concretely, we apply a priority-based label search on the GT mask. The procedure first checks whether the highest-priority label is present in the volume and, if so, selects the slice where that label occupies the largest area; if the label is absent, it proceeds to the next label in the priority list.The priority order is defined as follows, with label colors consistent with the GT visualization in Fig. 1:

1) *Non-enhancing Tumor Core (NETC)*, encoded as label 1 and shown in red, which typically corresponds to necrotic or cystic components within the tumor.
2) *Enhancing Tumor (ET)*, encoded as label 3 and shown in blue, which reflects actively enhancing regions that are clinically critical for assessing aggressiveness.
3) *Edema*, encoded as label 2 and shown in green, which often forms a broad, infiltrative region around the core tumor mass. This ordering is designed to favor slices that contain diagnostically salient subregions while avoiding an initialization rule that unintentionally benefits only a subset of modalities.

Placing edema at the lowest priority is particularly important for a modality-wise evaluation, because edema visibility depends strongly on imaging contrast, with FLAIR and T2w generally providing clearer edema delineation than T1-based modalities. If edema were prioritized, the starting slice would be systematically biased toward slices whose boundary cues are most prominent in T2w and FLAIR, potentially inflating apparent performance for those modalities during the initial prompting stage. In contrast, prioritizing NETC and ET yields a more balanced starting point that remains meaningful across modalities, since these subregions retain more consistent structural or enhancement-related signatures even when edema contrast is weak. As a result, the initialization strategy supports a fairer comparison of prompting performance across MRI modalities by limiting modality-specific advantages at the very first interaction step.

B. Generate Annotation

We evaluate eight prompting strategies based on sparse point prompts and bounding box prompts to characterize how different user interactions influence SAM2's zero-shot behavior in volumetric tumor delineation. Positive points are used to indicate tumor presence and provide an explicit foreground cue, whereas negative points serve as background constraints that discourage mask leakage into non-tumor tissue, which is particularly important in brain MRI where intensity overlap between tumor subregions and normal structures can be substantial. To keep the interaction budget comparable across strategies and to control for the stabilizing effect of background guidance, all point-based modes include a fixed pair of negative points.

The placement of negative points is designed to match the information available under each prompt type and to avoid inadvertently biasing one family of prompts. In modes 1–3, where only points are provided, the negative points are randomly sampled outside the tumor region to represent a realistic interaction pattern in which a user clicks a few background locations to prevent over-segmentation without relying on an explicit object extent. In contrast, in modes 5–7, a bounding box is already given and therefore defines a coarse spatial context for the target; accordingly, negative points are sampled inside the bounding box but outside the tumor region so that they explicitly suppress false positives within the region most likely to confuse the model, such as nearby edema-like intensities or partial-volume boundaries. Finally, mode 8 is the best-case prompt where we feed the predictor with the original ground truth mask to see the best possible segmentation performance from SAM2.

Although the ground truth annotation is possible to use, it is not directly aligned with real-world clinical setting where GT mask is not instantly available after acquiring the brain MRI. Hence, we only use that as reference and not referring as part of our evaluation. Fig. 4 visualizes the resulting annotations for each mode and highlights the systematic differences in negative-point placement across the point-only and box-assisted settings. The complete set of prompt modes is defined as follows:

1) *PN*: one positive point and a pair of negative points.
2) *2PN*: two positive points and a pair of negative points.
3) *3PN*: three positive points and a pair of negative points.
4) *B*: one bounding box surrounding the tumor object.
5) *BPN*: one bounding box, one positive point, and a pair of negative points.
6) *B2PN*: one bounding box, two positive points, and a pair of negative points.
7) *B3PN*: one bounding box, three positive points, and a pair of negative points.
8) *GT*: benchmark evaluation against the ground truth mask (upper-bound reference).

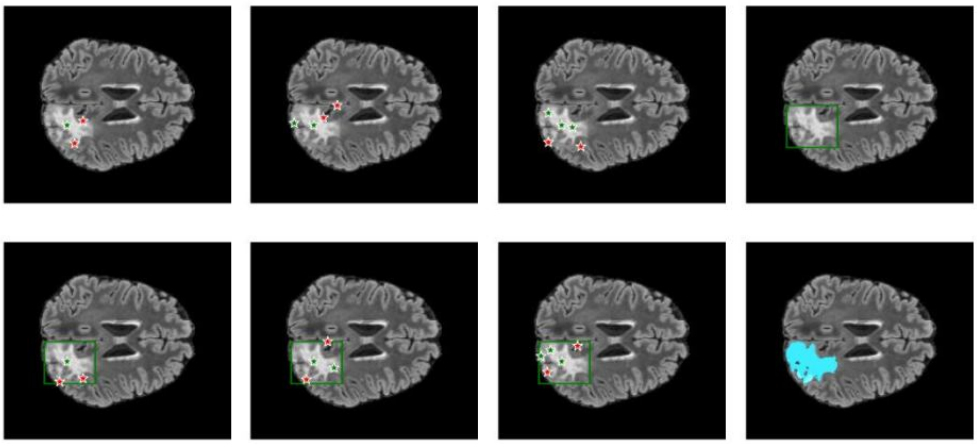


Fig. 4 The example of eight prompt annotation modes in our SAM2 evaluation.

C. Predict Initial Mask

Using the previously defined initialization slice and prompt annotations, we first query the SAM2 predictor to obtain an initial 2D segmentation on that single slice. This step establishes a slice-level mask that reflects how the model interprets the provided prompts before any volumetric propagation is performed. In practice, SAM2 returns multiple candidate masks for the same prompt, representing alternative segmentations with different confidence or granularity; for consistency across all experiments, we select the first mask in the returned list as the initial prediction. This initial 2D mask, together with its associated features, is then written into SAM2's memory mechanism, which serves as the reference for subsequent slice-to-slice propagation when predicting the segmentation across the remaining volume.

D. Mask Propagation

Following the bidirectional strategy in [20], we perform volumetric mask prediction in two propagation passes to reduce slice-to-slice drift that can arise when the model relies on a growing history of previous predictions. Starting from the initialization slice, we first run forward propagation, where SAM2 uses the initial 2D mask and its internal memory to infer masks on subsequent slices until reaching the last slice of the volume. This forward pass is effective for covering the portion of the tumor that extends in the positive slice direction, but a long unidirectional rollout can gradually accumulate small boundary errors, especially when tumor appearance changes abruptly across slices or when the target becomes faint.

After completing the forward pass, we reset the SAM2 state to clear the propagated memory and avoid carrying forward-pass biases into the opposite direction. We then reapply the same initialization prompts on the same starting slice to regenerate the initial 2D mask, ensuring that the backward traversal begins from an identical and well-defined reference. Using this refreshed initialization, we run backward propagation from the starting slice toward the first slice of the volume, which allows the model to track tumor extent in the negative slice direction under a clean memory context. The final volumetric prediction is obtained by combining the two passes, where slices after the starting slice are taken from the forward rollout and slices before the starting slice are taken from the backward rollout, with the starting slice anchored by the regenerated initial mask. An example of the resulting slice-wise propagation behavior is shown in Fig. 5, where the predicted tumor region (orange overlay) remains spatially coherent across neighboring slices while adapting to the gradual changes in tumor shape and size.

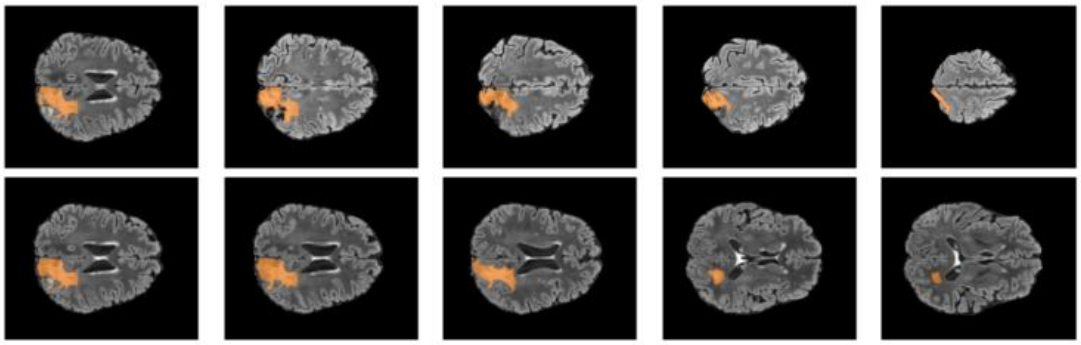


Fig. 5 Predicted mask after propagation is done. Top row shows the forward propagation result and the bottom row shows the backward propagation result with stride per frame set to 10

IV. EXPERIMENTAL SETUP

A. Dataset and Pre-processing

We selected 50 patient cases from the BraTS2024 Glioma Post-treatment dataset, a widely used benchmark for evaluating brain tumor segmentation methods [22]. The dataset is curated from clinically acquired MRI examinations and provided with expert-annotated ground truth (GT) labels, which supports reliable quantitative evaluation under heterogeneous real-world imaging conditions. To reduce selection bias and to ensure that performance reflects typical case variability rather than curated easy examples, we randomly sampled cases from the available cohort rather than filtering by tumor size, morphology, or apparent boundary clarity. Each case contains four MRI modalities, including native T1-weighted (T1n), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2w), and fluid-attenuated inversion recovery (FLAIR), together with the corresponding GT mask. Since SAM2 operates on 2D inputs in our pipeline, each modality volume is processed slice-wise during prompting and propagation; therefore, the 50 cases correspond to 200 modality volumes for evaluation.

In preprocessing, we intentionally keep the pipeline lightweight to better reflect a practical, training-free setting: we apply only pixel intensity normalization to each input slice and do not perform data augmentation, modality-specific enhancement, or task-driven preprocessing. This design choice avoids introducing additional learned priors or synthetic variability that could confound a zero-shot analysis, allowing the reported differences to be primarily attributed to modality contrast, prompt strategy, and propagation behavior rather than to augmentation-dependent effects.

B. Evaluation Methods

We evaluate segmentation quality using the Intersection over Union (IoU) and the Dice–Sørensen Coefficient (DSC), as both metrics are widely adopted in medical image segmentation and provide complementary views of overlap accuracy. IoU penalizes both false positives and false negatives through the union term, whereas DSC places more emphasis on foreground agreement, which is often informative when tumor regions occupy a relatively small portion of the image. The formula of Iou and DSC is given below:

$$\mathrm{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|}$$

$$\mathrm{DSC}(P, G) = \frac{2|P \cap G|}{|P| + |G|}$$

where *P* and *G* are the predicted segmentation and the ground truth segmentation respectively.

Although the BraTS ground truth masks contain multiple tumor subregion labels, in this study we convert the annotations to a binary tumor mask (tumor versus background) for evaluation. This choice aligns with the primary goal of our zero-shot analysis, which is to assess SAM2's ability to recover the overall tumor extent and maintain spatial consistency during slice-to-slice propagation, rather than to resolve fine-grained subregion boundaries that may require modality-specific cues and stronger task priors. Using a binary formulation also reduces ambiguity introduced by subregion label transitions and mitigates class imbalance effects that can dominate multi-class scores, thereby enabling a clearer comparison across MRI modalities and prompting strategies under a unified evaluation target. To compare segmentation performance across MRI modalities, we conduct paired statistical testing using a t-test, where the t-statistic measures the magnitude of the difference between modality-specific results relative to the variability across cases. We report the corresponding p-value, which quantifies the probability of observing a difference of the same magnitude under the null hypothesis of no true performance difference. Following common practice, p-values below 0.05 are interpreted as evidence that the observed modality-wise differences are unlikely to be explained by random variation alone.

V. RESULTS AND DISCUSSION

A. Evaluation on Prompt Perspective

Table 1 summarizes the average performance of the eight prompt modes, revealing a clear hierarchy between sparse point prompting and box-guided interaction. Point-only prompts (PN, 2PN, 3PN) produce limited overlap, with IoU in the range of 0.155–0.180 and Dice in the range of 0.228–0.262, suggesting that a few clicks, even when accompanied by consistent negative points, often provide insufficient spatial context to anchor the tumor extent in brain MRI where boundaries are weak and the foreground can be fragmented. Introducing a bounding box changes this behavior substantially, for example, the box-only setting (B) raises IoU to 0.355 and Dice to 0.470, indicating that coarse object extent is a strong prior for SAM2 in this task. The most reliable results emerge when the box is complemented with positive points (BPN, B2PN, B3PN), where the positive points act as targeted confirmations inside the tumor region while the box constrains the search space, reducing leakage and stabilizing ambiguous borders. This trend is reflected by the progressive improvements from BPN to B3PN, with B3PN achieving the best overall overlap (IoU 0.376, Dice 0.493) and B2PN consistently ranking second-best, while the GT entry serves as an upper-bound reference rather than a directly comparable prompt mode.

We further seek the evidence from the box plot distributions of IoU and Dice across MRI modalities and prompt modes, which provide a clearer view of stability beyond mean scores. As shown in Fig. 6, point-only prompts (PN, 2PN, 3PN) consistently yield low medians with wide interquartile ranges across all modalities, indicating that sparse clicks alone often fail to constrain the tumor extent and are highly sensitive to local intensity ambiguities. This behavior is especially apparent in the T1-based modalities, where the distributions concentrate near lower overlap values and exhibit frequent outliers, suggesting recurring under-segmentation or leakage when boundary cues are weak.

In contrast, the box prompt (B) shifts the distributions upward for both IoU and Dice, with noticeably higher medians and more compact spreads, reflecting that providing coarse object extent helps SAM2 localize the target more reliably even when modality contrast is suboptimal. The advantage becomes more pronounced when the bounding box is complemented with positive points (BPN, B2PN, B3PN). Across modalities, these combined strategies not only increase the central tendency but also reduce dispersion, implying improved robustness across heterogeneous cases. The positive points appear to act as disambiguating anchors within the boxed region, preventing the predictor from drifting toward adjacent tissues that share similar intensity patterns, while the box limits the search space and suppresses large-scale false positives. Notably, the strongest modality, T2-FLAIR, shows the highest and most stable distributions under box-assisted prompts, whereas T1ce and T1n still benefit substantially yet retain heavier tails, reflecting residual difficulty in cases where enhancement is sparse or non-enhancing components dominate. Overall, the box plot evidence reinforces that prompt design and MRI modality jointly shape both accuracy and reliability, and it motivates using box-assisted prompting as the more dependable initialization for subsequent slice-to-slice propagation in the 3D setting.

Table 1 Prompt modes average result. Point prompts (PN, 2PN, 3PN) are inferior compared to the box prompt. But the combination between box and point prompt (BPN, B2PN, B3PN) gives even better result. The GT result only serves for reference, not included in the comparison. Bold indicates best and underline indicates the second-best

Metric	Prompt Mode							
	PN	2PN	3PN	B	BPN	B2PN	B3PN	GT
IoU	0.155	0.180	0.167	0.355	<u>0.367</u>	0.374	0.376	0.494
Dice	0.228	0.262	0.247	0.470	0.482	<u>0.490</u>	0.493	0.601

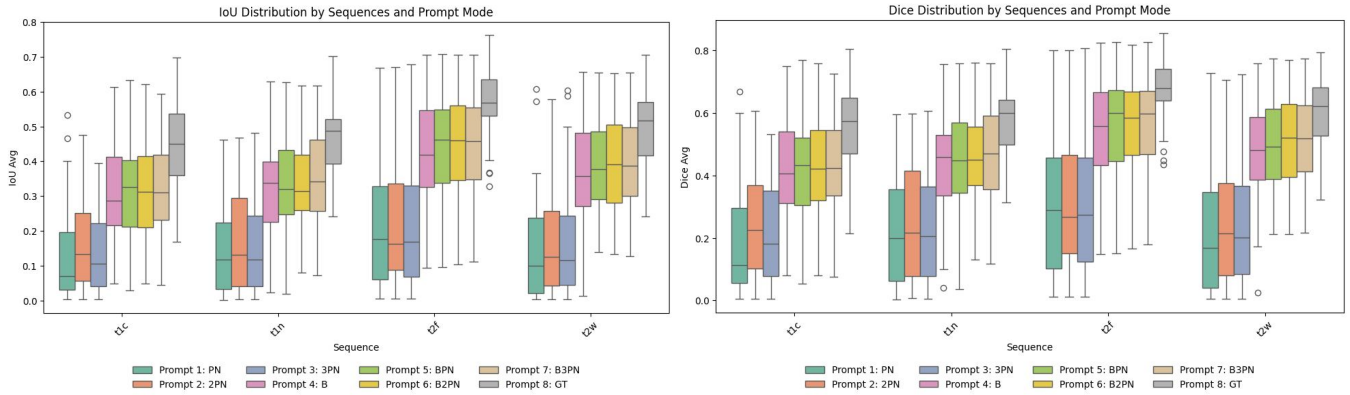


Fig. 6 Box plots of IoU (left) and Dice (right) for SAM2 zero-shot tumor segmentation across MRI modalities (T1ce, T1n, T2F, T2w) and prompt modes (PN/2PN/3PN, B, BPN/B2PN/B3PN, and GT). Results indicate that segmentation quality depends jointly on modality and prompting strategy, with box-assisted prompts (BPN–B3PN) generally yielding higher and more stable overlap than point-only prompts, and T2F/T2w providing stronger boundary cues than T1-based modalities; GT serves as an upper-bound reference.

B. Evaluation on Modality Perspective

The modality-wise analysis in Table 2 and the accompanying bar plots further contextualize these prompt effects by showing that overlap accuracy also depends on the visual contrast provided by each MRI modality, with T2-FLAIR achieving the highest average performance (IoU 0.374, Dice 0.480) compared with T1ce and T1n, which exhibit lower mean overlap. Importantly, this modality dependence does not contradict the prompt findings; instead, it explains why point-only prompts are more vulnerable to failure when boundary cues are subtle, while box-assisted prompts remain comparatively robust because they supply geometric guidance that is less sensitive to contrast variations. The error bars in the bar plots highlight substantial case-to-case variability across modalities, reinforcing that prompt choice is not merely a decorative detail but a practical control lever for reducing uncertainty, especially before moving from a reliable initial 2D mask to volumetric propagation in the subsequent evaluation.

Beyond the aggregated averages, the ordering in Table 2 suggests a clinically intuitive pattern: modalities that better expose diffuse abnormal signal yield stronger zero-shot delineation of the overall tumor extent under a binary target.

FLAIR ranks first, followed by T2w, while T1ce and T1n trail behind, which is consistent with the fact that a binary tumor mask implicitly rewards coverage of both core and peripheral abnormal regions. In many post-treatment cases, enhancing components can be sparse, heterogeneous, or even absent, and non-enhancing regions may blend into surrounding tissue, making T1-based boundaries less reliable for prompt-conditioned segmentation when the target is defined as tumor versus background. Conversely, T2-FLAIR and T2w tend to provide broader contrast for lesion extent and surrounding abnormality, so the model receives clearer spatial evidence once the initial prompt anchors the region, which helps explain why their mean overlap is higher and why their distributions remain elevated despite large case-to-case variation, as presented in Fig. 7.

Table 3 further supports this modality effect through pairwise statistical testing. The difference between T1ce and T1n is not statistically significant ($p = 0.146$), indicating that both T1-based modalities behave similarly under the evaluated prompting and propagation protocol. In contrast, comparisons between T2-FLAIR and each T1-based modality are strongly significant (T1ce vs T2-FLAIR: $t = -8.280$, $p < 0.001$; T1n vs T2-FLAIR: $t = -6.973$, $p < 0.001$), and T2w also significantly outperforms T1-based modalities (T1ce vs T2w: $t = -3.951$, $p < 0.001$; T1n vs T2w: $t = -2.574$, $p = 0.010$). Importantly, T2-FLAIR remains significantly stronger than T2w ($t = 4.332$, $p < 0.001$), reinforcing that the suppression of free-fluid signal in FLAIR can sharpen lesion-related hyperintensity and make the tumor extent more consistently separable from background tissue. Taken together, these results suggest that when SAM2 is applied in a zero-shot, prompt-driven workflow for binary tumor delineation, modality selection is not merely a presentation choice but a primary determinant of achievable overlap, with T2-FLAIR offering the most favorable contrast for stable volumetric segmentation in this setting.

Table 2 The overall result of each sequence. These numbers are derived from the average of IoU and Dice of each prompt from mode 1 to mode 8. Bold indicates best performance and underline indicates second-best performance

Sequence	IoU Avg	Dice Avg
T1c	0.264	0.359
T1n	0.281	0.383
FLAIR	0.374	0.480
T2	<u>0.314</u>	<u>0.415</u>

Table 1 T-Test result of all MRI sequences. Larger T value and P closer to 0 indicates a stronger difference between the two subjects.

Comparison	T-Statistic	P-Value
T1ce vs T1n	-1.454	0.146
T1ce vs T2f	-8.280	0.000
T1ce vs T2w	-3.951	0.000
T1n vs T2f	-6.973	0.000
T1n vs T2w	-2.574	0.010
T2f vs T2w	4.332	0.000

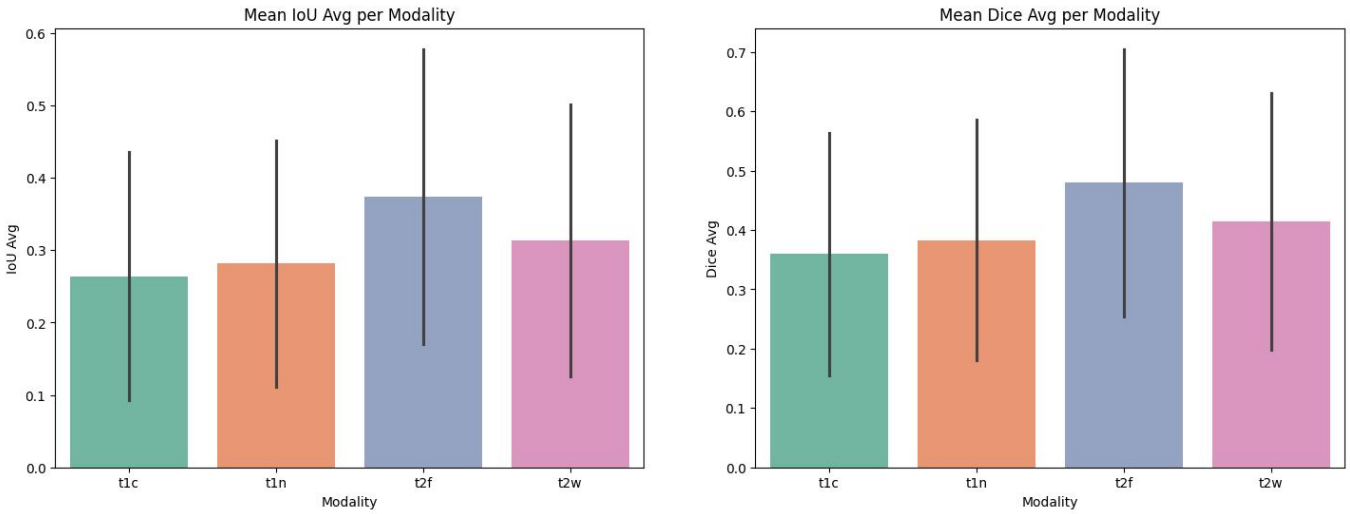


Fig. 7 Mean overlap performance of SAM2 zero-shot tumor segmentation across MRI modalities, shown as average IoU (left) and average Dice (right) with variability indicated by error bars. The modality-wise trends suggest that T2-FLAIR (T2F) provides the most informative contrast for delineating overall tumor extent, while T1ce/T1n exhibit lower average overlap, reflecting weaker or more ambiguous boundary cues for prompt-driven segmentation.

C. Limitations

A key limitation of this study is that evaluation is performed using a binary tumor mask, obtained by collapsing the original multi-class BraTS annotations into tumor versus background. While this choice enables a clean and consistent assessment of zero-shot delineation and propagation stability, it can also introduce a modality-dependent bias because modalities such as T2-FLAIR and T2w typically highlight broader lesion extent, whereas T1-based modalities may emphasize different subregions and therefore be disadvantaged when the target rewards overall coverage. Moreover, binary scoring does not reveal whether SAM2 can separate clinically meaningful subregions, such as enhancing tumor versus non-enhancing core, which remains an important requirement for comprehensive tumor characterization. Nevertheless, the study remains for its intended scope: the binary formulation isolates the central question of whether a prompt-driven, training-free pipeline can reliably recover global tumor extent and maintain slice-to-slice consistency across modalities, which is a practical prerequisite for any interactive volumetric workflow and a meaningful first step before extending the analysis to multi-class delineation and subregion-specific prompting.

VI. CONCLUSION

In this work, we investigated SAM2 as a training-free, prompt-driven alternative for volumetric brain tumor delineation on BraTS2024 post-treatment MRI, with an emphasis on how prompting design and MRI modality jointly shape zero-shot segmentation quality and propagation stability.

By standardizing the initialization slice selection and evaluating eight prompt modes, we observed a consistent advantage for box-guided prompting, where a bounding box provides reliable coarse extent and additional positive points further stabilize the prediction, yielding higher IoU and Dice than point-only strategies. The distributional evidence from box plots reinforces that this improvement is not limited to mean gains, but also reflects reduced variability and fewer unstable cases, which is critical when the initial 2D mask becomes the anchor for volumetric propagation.

From the modality perspective, the results show that SAM2's overlap accuracy is strongly influenced by modality-specific contrast, with FLAIR achieving the highest average performance, followed by T2w, while T1ce and T1n remain comparatively lower and statistically similar. Pairwise t-tests further substantiate these differences, indicating that the performance gaps between TFLAIR and the other modalities are unlikely to be explained by random variation. Taken together, these findings suggest that effective zero-shot volumetric tumor segmentation with SAM2 benefits from combining geometric guidance through box-assisted prompts with modalities that provide clearer lesion extent cues, while acknowledging that our binary evaluation focuses on global tumor extent rather than subregion separation.

REFERENCES

[1] Buchner, J.A., Peeken, J.C., Etzel, L., Ezhov, I., Mayinger, M., Christ, S.M., Brunner, T.B., Wittig, A., Menze, B.H., Zimmer, C., Meyer, B., Guckenberger, M., Andratschke, N., El Shafie, R.A., Debus, J., Rogers, S., Riesterer, O., Schulze, K., Feldmann, H.J., Blanck, O., Zamboglou, C., Ferentinos, K., Bilger, A., Grosu, A.L., Wolff, R., Kirschke, J.S., Eitz, K.A., Combs, S.E., Bernhardt, D., Rueckert, D., Piraud, M., Wiestler, B., Kofler, F.: Identi-fying core MRI sequences for reliable automatic brain metastasis segmentation. Radiothera-py and Oncology, 188, 109901 (2023). https://doi.org/10.1016/j.radonc.2023.109901

[2] Gaillard, F., Baba, Y., Bell, D., et al.: MRI sequences (overview). Radiopaedia.org (2025). https://doi.org/10.53347/rID-37346

[3] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment Anything in Images and Videos. arXiv pre-print (2024). https://arxiv.org/abs/2408.00714

[4] Walsh, J., Othmani, A., Jain, M., Dev, S.: Using U-Net network for efficient brain tumor segmentation in MRI images. Healthcare Analytics, 2, 100098 (2022). https://doi.org/10.1016/j.health.2022.100098

[5] Lin, S.Y., Lin, C.L.: Brain tumor segmentation using U-Net in conjunction with Efficient-Net. PeerJ Comput Sci, 10, e1754 (2024). https://doi.org/10.7717/peerj-cs.1754

[6] Agrawal, P., Katal, N., Hooda, N.: Segmentation and classification of brain tumor using 3D-UNet deep neural networks. International Journal of Cognitive Computing in Engineering, 3, 199–210 (2022). https://doi.org/10.1016/j.ijcce.2022.11.001

[7] Isensee, F., Jaeger, P.F., Kohl, S.A.A., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods, 18, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z

[8] Shen, C., Fang, Y., Yu, X., Guo, C., Ju, Z.: CS-UNet: A LightWeight UNet Model Based On Context Information. In: Lan, X., Mei, X., Jiang, C., Zhao, F., Tian, Z. (eds.) Intelligent Robotics and Applications, ICIRA 2024. LNCS, vol. 15205. Springer, Singapore (2025). https://doi.org/10.1007/978-981-96-0777-8_24

[9] Wang, P., Yang, Q., He, Z., Yuan, Y.: Vision transformers in multi-modal brain tumor MRI segmentation: A review. Meta-Radiology, 1(1), 100004 (2023). https://doi.org/10.1016/j.metrad.2023.100004

[10] Zhou, T.: M2GCNet: Multi-Modal Graph Convolution Network for Precise Brain Tumor Segmentation Across Multiple MRI Sequences. IEEE Transactions on Image Processing, 33, 4896-4910 (2024). https://doi.org/10.1109/TIP.2024.3451936

[11] Zhu, Z., He, X., Qi, G., Li, Y., Cong, B., Liu, Y.: Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. Information Fusion, 91, 376-387 (2023). https://doi.org/10.1016/j.inffus.2022.10.022

[12] Zhang, G., Zhou, J., He, G., Zhu, H.: Deep fusion of multi-modal features for brain tumor image segmentation. Heliyon, 9(8), e19266 (2023). https://doi.org/10.1016/j.heliyon.2023.e19266

[13] Kirillov, A., et al.: Segment Anything. 2023 IEEE/CVF International Conference on Com-puter Vision (ICCV), Paris, France, 3992-4003 (2023). https://doi.org/10.1109/ICCV51070.2023.00371

[14] Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Med Image Anal, 89, 102918 (2023). https://doi.org/10.1016/j.media.2023.102918

[15] Mattjie, C., et al.: Zero-Shot Performance of the Segment Anything Model (SAM) in 2D Medical Imaging: A Comprehensive Evaluation and Practical Guidelines. 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), 108-112 (2023). https://doi.org/10.1109/BIBE60311.2023.00025

[16] Zhang, L., Deng, X., Lu, Y.: Segment Anything Model (SAM) for Medical Image Segmen-tation: A Preliminary Review. 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 4187-4194 (2023). https://doi.org/10.1109/BIBM58861.2023.10386032

[17] Ali, L., Alnajjar, F., Swavaf, M., Elharrouss, O., Abd-Alrazaq, A., Damseh, R.: Evaluating segment anything model (SAM) on MRI scans of brain tumors. Scientific Reports, 14(1), 21659 (2024). https://doi.org/10.1038/s41598-024-72342-x

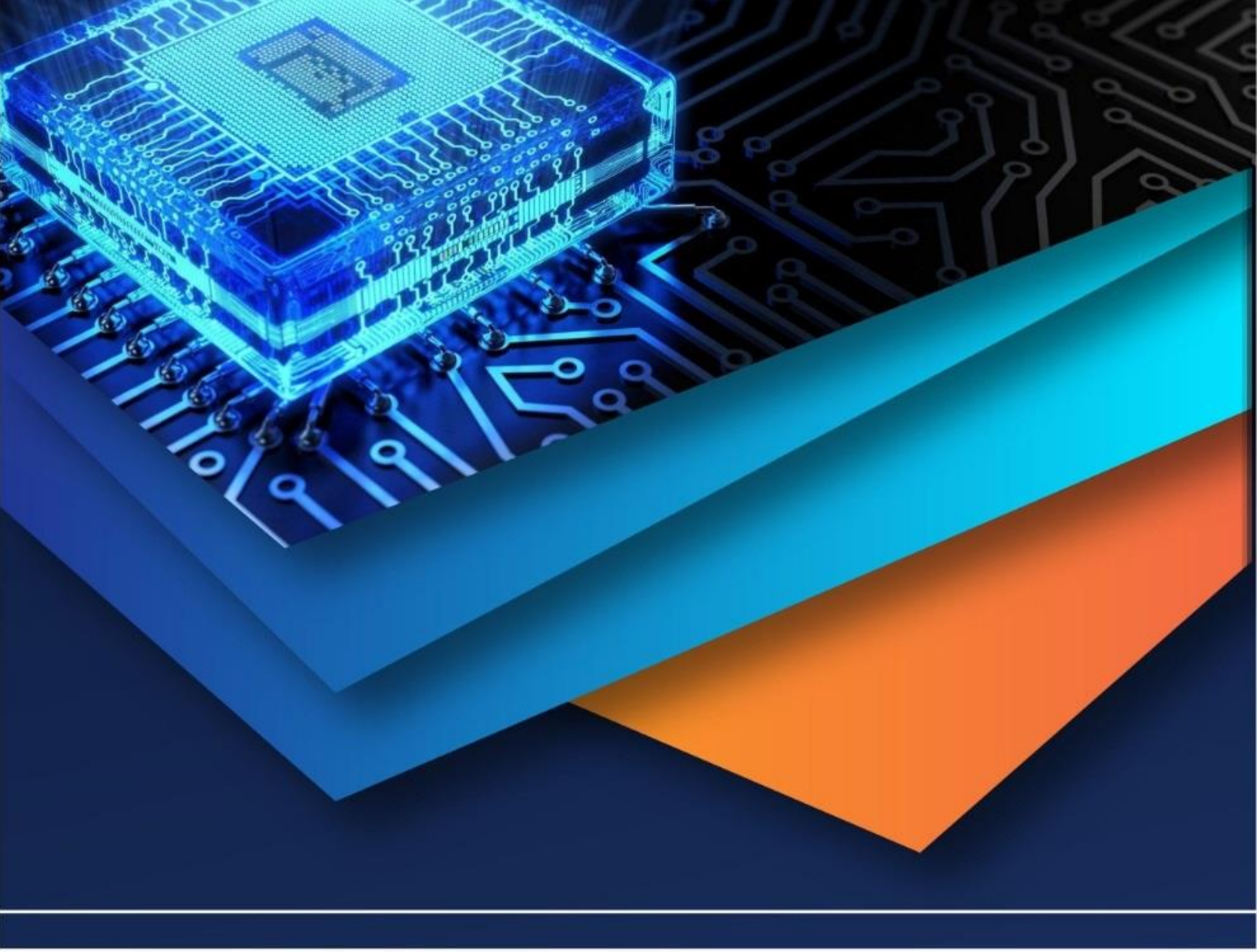
[18] Yang, Y., Wu, X., He, T., Zhao, H., Liu, X.: SAM3D: Segment Anything in 3D Scenes. arXiv preprint (2023). https://arxiv.org/abs/2306.03908

[19] Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: SAM-Med3D: Towards General-purpose Segmenta-tion Models for Volumetric Medical Images. arXiv preprint (2024). https://arxiv.org/abs/2310.15161

[20] Dong, H., Gu, H., Chen, Y., Yang, J., Chen, Y., Mazurowski, M.A.: Segment Anything Model 2: An Application to 2D and 3D Medical Images. arXiv preprint (2024). https://arxiv.org/abs/2408.00756

[21] Zhao, X., et al.: Inspiring the Next Generation of Segment Anything Models: Comprehen-sively Evaluate SAM and SAM 2 with Diverse Prompts Towards Context-Dependent Con-cepts under Different Scenes. arXiv preprint (2024). https://arxiv.org/abs/2412.01240

[22] Correia de Verdier, M., et al.: The 2024 Brain Tumor Segmentation (BraTS) Challenge: Gli-oma Segmentation on Post-treatment MRI. arXiv preprint (2024). https://arxiv.org/abs/2405.18368



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY