Evaluating the Adaptation Potential of SAM2 for Glacier Segmentation in severe Weather

Bindusara Nagathihalli Lokesh [1], Laura Camila Duran Vergara [1], Hans-Gerd Maas [1], Anette Eltner [1]

[1] Institute of Photogrammetry and Remote Sensing, Dresden University of Technology, Germany -
(bindusara.nagathihalli_lokesh, laura_camila.duran_vergara, hans-gerd.maas, anette.eltner) @tu-dresden.de

Keywords: Glacier segmentation, Foundation Model, Segment Anything Model 2 (SAM2), Automation pipeline, Mask quality estimation

Abstract

Ground based time lapse cameras provide continuous, high frequency observations of glacier dynamics; however, automated analysis of these image streams remains challenging due to fog, snowfall, lens contamination, and variable illumination. This study investigates the potential of adapting the foundation segmentation model Segment Anything Model 2 (SAM2) for glacier segmentation from ground-based monitoring. To enable integration into automated pipelines, SAM2 is configured in image mode with a learned prompt generation strategy, while fine-tuning is restricted to the prompt encoder and mask decoder. In addition, the internal Intersection over Union (IoU) prediction head is utilized as a confidence estimator to assess segmentation reliability. Experimental results demonstrate that the adapted model achieves stable segmentation under moderate environmental variability, while degrading under severe visibility loss. This stability is consistent across model scales and input resolutions. The confidence estimation further provides a meaningful signal for identifying uncertain predictions, supporting reliability-aware processing in downstream workflows.

1. Introduction

Ground based photogrammetric monitoring systems provide high spatio-temporal resolution observations of geomorphological and cryospheric processes.(Grothum et al., 2025) In glacier environments, monoscopic time-lapse cameras enable continuous documentation of frontal dynamics and surface evolution (Schwalbe and Maas, 2017). These image sequences form the foundation of long term monitoring pipelines, where they are used for motion estimation, surface reconstruction, and change analysis (James et al., 2017)

Within such pipelines, segmentation serves as a critical preprocessing step by separating relevant scene components, such as glacier fronts, water bodies, and surrounding terrain. The quality of this preprocessing stage directly influences downstream tasks (Eltner et al., 2016), as errors in segmentation can propagate into downstream reconstruction and measurement processes, a dependency observed in related settings such as motion segmentation for Structure from Motion pipelines (Goli et al., 2024) and photogrammetric change detection workflows (Ulm et al., 2025). Consequently, segmentation in this context must not only be accurate but also stable and reliable across extended temporal sequences.

Achieving stable segmentation, however, is particularly challenging in ground-based environmental monitoring scenarios. Unlike controlled datasets commonly used in computer vision, these terrestrial systems operate under highly variable and uncontrolled acquisition conditions. Variations in illumination, atmospheric visibility, snowfall, fog, and lens contamination introduce significant variability into the image data. These factors result in non-stationary data distributions and frequent deviations from conditions represented in training datasets, ultimately leading to degraded segmentation performance over time (Zhu et al., 2017).

These challenges are illustrated in Figure 1, where environmental disturbances such as lens contamination and atmospheric effects significantly alter scene visibility and structure. Such degradations introduce ambiguous boundaries, partial occlusions, and reduced contrast, making consistent segmentation increasingly difficult.

Figure 1. Examples of environmental variability in glacier monitoring: lens droplets causing local distortion (left) and reduced visibility due to fog or snow (right).

Given these challenges, robustness alone is insufficient for reliable deployment in automated monitoring pipelines. Since segmentation outputs serve as inputs for further processing (Duran Vergara et al., 2026), it is essential not only to maintain performance under varying conditions, but also to identify cases where predictions may be unreliable due to environmental degradation or distribution shifts. This necessitates mechanisms for estimating prediction confidence alongside segmentation. The central challenge, therefore, is to develop approaches that can provide both stable segmentation under moderate variability and reliable indicators of prediction quality.

This study investigates the adaptation of a foundation segmentation model for ground-based glacier monitoring under challenging environmental conditions. Rather than focusing solely on segmentation accuracy, we analyze the robustness of the model across varying conditions, input resolutions, and model configurations, and examine its ability to estimate prediction reliability through an internal confidence mechanism. The goal is to assess the feasibility of integrating such models into long term monitoring pipelines, where both segmentation stability and prediction reliability are essential.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-205-2026

2. Related Work

Automated segmentation plays a key role in environmental monitoring workflows, where it is used to extract relevant scene components. For example, camera-based river water level estimation relies on robust water surface segmentation despite changes in illumination and weather conditions (Blanch et al., 2026). Similarly, segmentation networks such as SAU-Net (Xiang et al., 2025) have been applied for glacier and ice feature delineation in remote sensing imagery, highlighting the importance of accurate boundary extraction in cryospheric studies.

These and related applications are typically based on convolutional or transformer-based segmentation models. While convolutional architectures (Xiang et al., 2025) provide strong local feature extraction, transformer-based models such as the Swin Transformer (Filho et al., 2025) and its adaptations to remote sensing tasks (Meena et al., 2025) introduce mechanisms for capturing global spatial context through self-attention. Comparative studies across a wide range of convolutional neural network architectures have shown that segmentation performance in environmental imagery is strongly influenced by augmentation strategies and domain-specific variability (Wagner et al., 2023).

However, both convolutional neural networks and transformer based model families rely on sufficient coverage of environmental variability within the training data and often require extensive annotated datasets to achieve robust performance across diverse and evolving conditions. In practice, this dependence limits their ability to generalize to environmental image streams characterized by non-stationary data distributions, seasonal variations, and atmospheric disturbances (Zhu et al., 2017).

This limitation motivates the exploration of more generalizable segmentation approaches that are less dependent on task specific training data. More recently, foundation segmentation models have emerged as a promising alternative for improving cross domain generalization. The Segment Anything Model (SAM) (Kirillov et al., 2023) introduced a prompt driven, class agnostic segmentation framework trained on large scale heterogeneous datasets. Its successor, SAM2, extends this paradigm to both image and video modalities and incorporates additional capabilities such as multimask prediction and mask quality estimation (Ravi et al., 2024).

Initial studies have explored the application of SAM and SAM2 in wide ranging environmental contexts. For example, SAM-based approaches have been applied to glaciological feature segmentation across multiple remote sensing platforms (Shankar and Veen, 2023), as well as in related domains such as underwater scene analysis (Lian and Li, 2024) and vegetation mapping through integration with LiDAR data (Zhu et al., 2025). These studies highlight the flexibility of foundation models in adapting to diverse data sources and segmentation tasks.

Despite these advances, the application of SAM2 in ground-based, long term environmental monitoring remains largely unexplored. In particular, challenges related to severe atmospheric variability, domain specific adaptation, and prediction reliability have not been systematically analyzed. Furthermore, while SAM2 provides an internal mechanism for mask quality estimation, its effectiveness as a proxy for segmentation reliability in real world monitoring pipelines has not been thoroughly evaluated.

These limitations motivate the present study, which focuses on analyzing the adaptation behavior of SAM2 under challenging environmental conditions and assessing its suitability for reliable deployment in automated monitoring systems.

3. Dataset and Annotation Strategy

The image dataset used in this study was collected within the AI4Glaciers project (Duran Vergara et al., 2026), which operates an autonomous Multi-View Stereo (MVS) system (Mallalieu et al., 2017) installed in front of the lake terminating Glacier Perito Moreno (GPM) in southern Patagonia, Argentina. The system consists of eight fixed RGB cameras that acquire two consecutive images at each 30 minute interval during daylight hours. The monitoring setup operates under challenging natural conditions that frequently affect image quality. These variations introduce changes in contrast, visibility, and local texture, which directly influence both segmentation reliability and photogrammetric reconstruction.

For segmentation purposes, five surface classes have been defined: Glacier, Water, Sky, Rocks, and Mountain. Among these, the glacier class is the most relevant for downstream 3D reconstruction and change detection, while rocks provide spatial context and stable reference regions(Blanch et al., 2021). The class definitions were chosen to reflect both photogrammetric relevance and visual separability in the field of view. In particular, accurate delineation of the glacier boundary is critical for subsequent surface reconstruction and change analysis.

3.1 Annotation Strategy

Given the temporal density of the dataset and the high visual similarity between consecutive frames, an efficient annotation strategy was required. A SAM-based annotation tool (Bhattiprolu, 2024) was employed to facilitate interactive mask generation, significantly reducing manual effort. This workflow was further enhanced by incorporating SAM2-based mask propagation (Lokesh et al., 2026), enabling rapid transfer of annotations across temporally adjacent frames and accelerating the overall labeling process.

3.2 Augmentation strategy

Although the time-lapse dataset captures a wide range of naturally occurring environmental conditions, long term glacier monitoring inevitably encounters additional visibility degradation and illumination variations that may not be sufficiently represented in the annotated subset. To improve robustness, an augmentation pipeline was designed to reflect typical environmental and optical disturbances observed in the deployed monitoring system.

The pipeline combines standard geometric and photometric augmentations (Wagner et al., 2023) with targeted visibility related effects. Geometric transformations (horizontal flipping, moderate scaling, and rotation) were applied consistently to images and masks to increase spatial variability. Photometric variations such as brightness contrast adjustments, gamma shifts, color jitter, noise injection, and mild blur were included to account for exposure fluctuations and sensor noise.

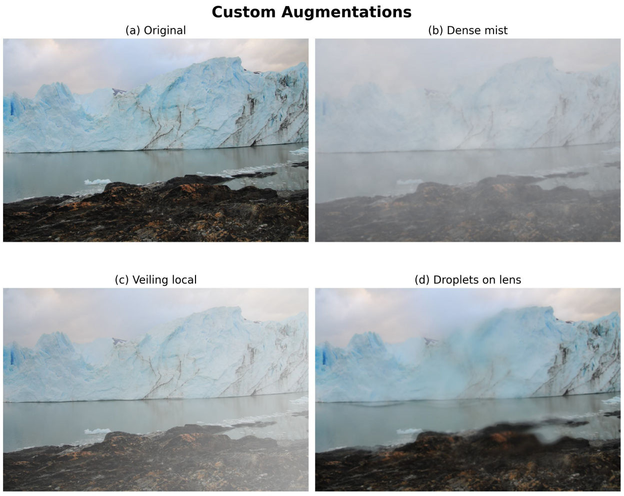


Figure 2. Custom environment oriented augmentations: (a) original image, (b) dense mist with reduced visibility and contrast, (c) local veiling simulating mist onset, and (d) droplet-based distortion simulating lens contamination.

To improve robustness beyond standard augmentation pipelines, the training data was enriched using environment oriented transformations designed to better reflect the optical and atmospheric conditions observed in the glacier monitoring setup. As illustrated in Fig. 2, dense mist was simulated through a combination of global contrast suppression, brightness adjustment, blur, resampling induced softening, and large scale fog like veiling (Fig. 2b), thereby reproducing the reduced visibility and loss of local detail observed in real sequences. Similarly, local veiling and glare effects (Fig. 2c) were used to mimic backlighting and diffuse light scattering, while droplet based augmentations (Fig. 2d) introduced localized distortion, blur, and partial occlusion patterns resembling water or moisture on the protective housing.

The resulting segmentation dataset was expanded not only in size but also in environmental variability, making it more representative of the severe outdoor conditions encountered during long term glacier monitoring. The objective of this augmentation strategy is not to synthetically reproduce all environmental processes, but to expose the model to plausible variations in visibility and optical distortion.

3.3 Dataset Construction

The original glacier images were captured at a resolution of 6000 × 4000 pixels. Since SAM2 (Ravi et al., 2024) operates on fixed size inputs (1024 × 1024), training samples were generated through a controlled patch extraction process rather than simple global resizing. Variable sized rectangular patches were sampled from the high resolution imagery, with constraints ensuring that each patch contained at least two to three distinct semantic classes. This gating strategy avoids trivial single class crops and promotes structurally meaningful training samples. The extracted patches were subsequently resized to the network input resolution and augmented using geometric and photometric transformations as described in the previous section.

Based on approximately 660 manually annotated images obtained through labeling and propagation, the dataset was expanded to yield around 2100 training samples, along with validation and test sets comprising approximately 700 images in total, following a 60:20:20 train–validation–test split.

To mitigate data leakage, temporally adjacent frames captured within the same acquisition interval were explicitly separated during dataset construction, ensuring that highly similar or near duplicate frames were not distributed across training and validation splits. This separation strategy reduces the risk of inflated performance due to near duplicate samples. However, given the fixed camera setup, some degree of structural similarity across splits may remain, as images acquired from the same viewpoint share consistent scene geometry and background context.While the camera viewpoint remains fixed, the dataset exhibits substantial variability in environmental conditions, including illumination, atmospheric visibility, and scene dynamics such as glacier motion and water level changes. Therefore, the reported performance reflects robustness to temporal and environmental variability, rather than spatial generalization to unseen viewpoints. Generalization to entirely new glacier sites remains outside the scope of this study.

4. Methodological Framework

4.1 Image Mode Adaptation of SAM2

This study investigates the adaptability of SAM2 (Ravi et al., 2024) for glacier monitoring under severe environmental variability. To enable fully automated segmentation, a prompt generation strategy inspired by (Xie et al., 2025) is adopted, as described in prior work (Lokesh et al., 2026). An adaptive prompt generator is used, as fixed or manual prompting is not suitable due to continuous scene changes across long temporal sequences.

SAM2 is built upon a hierarchical transformer-based image encoder (Hiera), pretrained on large scale segmentation datasets. In this work, the image encoder is kept frozen during fine-tuning, while adaptation is restricted to the prompt encoder and mask decoder. This setup preserves the general feature representations learned during pretraining, reduces computational complexity, and enables evaluation of whether these pretrained features are sufficient to support robust segmentation under challenging environmental conditions.

Figure 3 illustrates the adapted pipeline. The image encoder extracts multi-scale features, which are used by a learned prompt generator to produce class specific prompts. These prompts are encoded and passed to the mask decoder, which generates segmentation masks along with an associated confidence estimate.

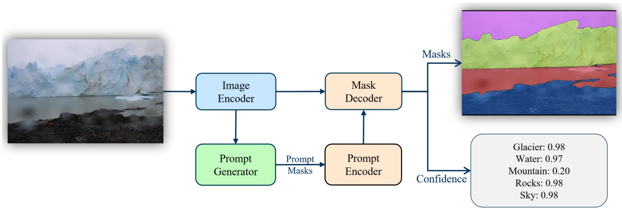


Figure 3. Overview of the adapted SAM2 pipeline for automated glacier segmentation, producing segmentation masks and associated confidence scores.

4.2 Joint Fine-Tuning and Confidence Supervision

In contrast to the original SAM2 (Ravi et al., 2024) configuration, which produces multiple candidate masks per prompt, the multimask output branch is disabled to enforce deterministic single mask predictions. Furthermore, to enable stable and reliable mask generation for the SfM-MVS workflow, the internal IoU prediction head, originally designed to rank multiple candidate masks, is repurposed as a reliability estimator for the generated segmentation.

During training, the predicted segmentation mask for each class is compared with the corresponding reference mask to compute the true IoU. The IoU prediction head is supervised using a regression loss to align predicted confidence with actual segmentation quality, and is optimized jointly with the segmentation loss.

4.3 Experimental Controls

To analyze the trade off between computational efficiency and segmentation robustness, controlled experiments were conducted under varying input resolutions, model variants, and finetuning strategies.

- **Resolution Variation:** The standard configuration uses 1024×1024 inputs, consistent with the default SAM2 (Ravi et al., 2024) setting. Additional experiments were performed at reduced and increased resolutions. When modifying the input resolution, only the spatial dimensions of the backbone feature maps were adjusted while preserving the hierarchical scaling structure of the Hiera encoder (e.g., feature resolutions such as /4, /8, /16 relative to input size). No architectural changes were introduced.
- **Model Variants:** Multiple SAM2 backbone variants were evaluated to assess the influence of encoder capacity on segmentation robustness.
- **Fine-Tuning Ablation Study:** To assess the extent of task-specific adaptation required, a small ablation study was conducted by progressively unfreezing components of the mask decoder, starting from the output hypernetworks and extending through intermediate convolutional layers to the final attention layer of the two-way transformer. In the initial configuration, the prompt encoder was kept fully frozen. This incremental setup allows evaluating whether performance gains arise from broad decoder adaptation or from a limited set of task specific calibration layers.

All experiments were conducted under identical optimization settings to ensure fair and controlled comparison

5. Results

5.1 Segmentation Performance under Clear Conditions

Quantitative evaluation of the adapted SAM2 (Ravi et al., 2024) model under clear weather conditions demonstrates strong segmentation performance, with a macro-averaged IoU of 0.95 and F1 score of 0.97 (Table 1).

High segmentation accuracy is observed across all foreground classes, indicating that the pretrained SAM2 encoder features, combined with lightweight decoder side adaptation, are sufficient to capture the structural characteristics of glacier scenes under favorable visibility conditions. The background class exhibits lower performance (IoU = 0.233), which is expected due to its definition. In this setup, boundary pixels between semantic regions are explicitly assigned to the background class to support flexibility in the SfM-MVS processing pipeline

Class ID	IoU	F1 Score	Precision	Recall
0 (background)	0.233	0.378	0.334	0.434
1 (glacier)	0.961	0.980	0.982	0.979
2 (mountain)	0.928	0.963	0.958	0.967
3 (rocks)	0.973	0.986	0.980	0.992
4 (sky)	0.966	0.983	0.987	0.978
5 (water)	0.927	0.962	0.977	0.948
Macro Avg	**0.951**	**0.975**	**0.977**	**0.973**

Table 1. Per class segmentation performance under clear weather conditions.

(Duran Vergara et al., 2026). As a result, the background class aggregates ambiguous boundary regions from multiple classes, leading to inherently lower separability and increased label uncertainty. Consequently, errors in this class primarily reflect boundary ambiguity rather than large scale misclassification. The confusion matrix (Fig. 4) further illustrates these patterns, showing strong diagonal dominance across all foreground classes, with off diagonal entries concentrated around boundary related transitions (e.g., glacier–water and glacier–mountain interfaces).

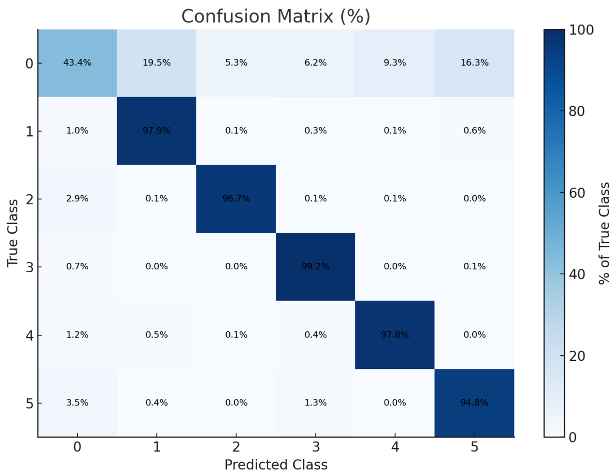


Figure 4. Confusion matrix for SAM2 segmentation results under clear weather conditions, showing strong class-wise prediction accuracy.

5.2 Segmentation Performance under Degraded Visibility Conditions

To assess robustness under adverse environmental variability, the adapted SAM2 model was evaluated under progressively challenging visibility conditions. These include moderate degradation scenarios, such as slight haze or mist and water droplets on the camera lens (as illustrated in Fig. 1), as well as severe degradation in the form of dense mist, where scene structure becomes significantly obscured (see Fig. 5).

While moderate conditions primarily introduce localized distortions and partial occlusions, severe visibility loss leads to a substantial reduction in contrast and structural detail, making consistent segmentation considerably more challenging.

Under slightly degraded visibility conditions, the model maintains strong performance, achieving a macro IoU of 0.86 and F1 score of 0.93. As shown in Fig. 6, the confusion matrix remains largely diagonally dominant, with errors primarily localized to regions affected by optical distortions.

Figure 5. Example of severe visibility degradation due to dense mist, resulting in significant loss of structural detail and contrast

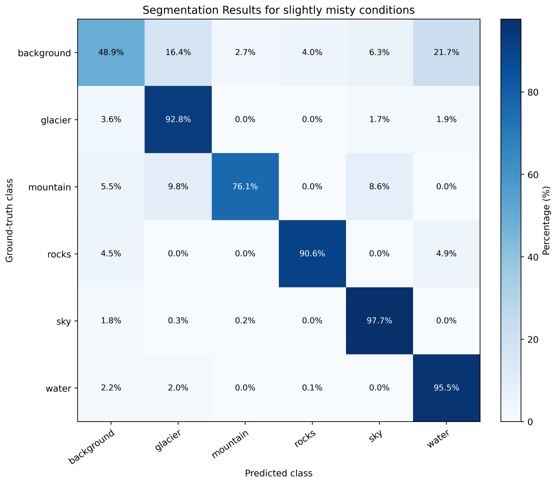


Figure 6. Confusion matrix under slightly degraded visibility conditions. High diagonal dominance indicates preserved class separability with minor boundary ambiguity.

In contrast, severe visibility degradation caused by dense mist leads to a substantial drop in performance, with macro IoU decreasing to 0.48 and F1 score to 0.62. As illustrated in Fig. 7, the confusion matrix shows increased off diagonal entries, reflecting the loss of structural and textural cues under heavy atmospheric occlusion.

5.3 Confidence Head Analysis and IoU Prediction Reliability

To further analyze the reliability of the adapted model, we evaluate the behavior of the confidence head, which predicts the IoU between the predicted and ground truth masks. This analysis provides insight into whether the model is able to internally estimate the quality of its own predictions.

Figure 8 shows the distribution of signed errors between predicted IoU and actual IoU across all classes. For all the foreground classes (glacier, mountain, rocks, sky, and water), the error distribution is centered around zero with a sharp peak, indicating well calibrated IoU estimation. In contrast, the background class exhibits a negative bias, with errors skewed toward underestimation. This behavior aligns with the definition of background regions at object boundaries, where ambiguity is inherently higher.

The relationship between predicted and actual IoU is further illustrated in Figure 9, which highlights where deviations occur

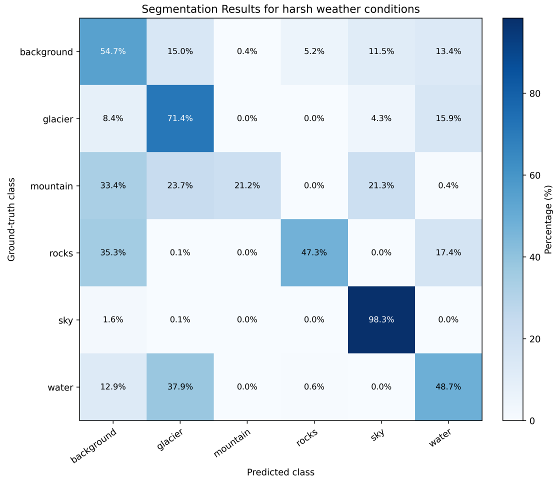


Figure 7. Confusion matrix under severe mist conditions. Reduced diagonal dominance and increased inter-class confusion reflect the impact of global visibility loss.

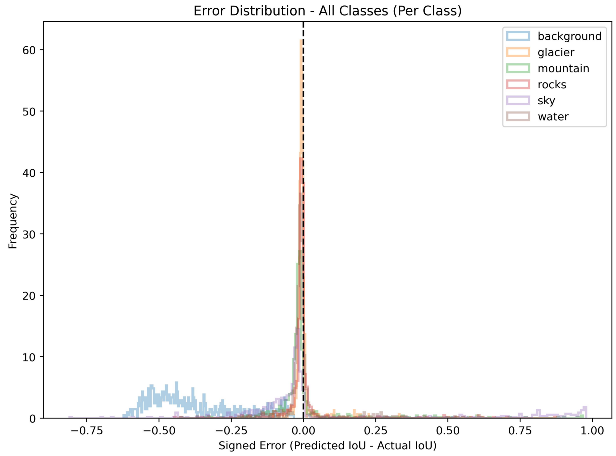


Figure 8. Distribution of signed error between predicted IoU and actual IoU across all classes. A zero centered distribution indicates well calibrated confidence estimates.

across different classes. Notably, increased dispersion is observed for sky and mountain classes, primarily due to cloud cover and atmospheric occlusions that partially or fully obscure structural boundaries, leading to ambiguity between these regions. Similarly, mild deviations are present in the rocks class, which can be attributed to visual similarity with large ice blocks and water patches that were intentionally grouped under a single class for consistency.

In contrast, the glacier class exhibits a tightly clustered distribution around the diagonal, indicating stable segmentation performance and reliable IoU estimation. As shown in Fig. 10, the error distribution is centered near zero (Fig. 10a), and predicted versus actual IoU values closely follow the diagonal (Fig. 10b), with most samples concentrated in the 0.95 to 1.00 range. This indicates that the majority of predictions are both highly accurate and assigned high confidence, with only minor deviations typically remaining above the 80% IoU range. This behavior is particularly relevant for photogrammetric reconstruction and SfM-MVS pipelines, where accurate and consistent delineation of scene components is critical (Ulm et al., 2025).

To complement the quantitative evaluation, qualitative seg-

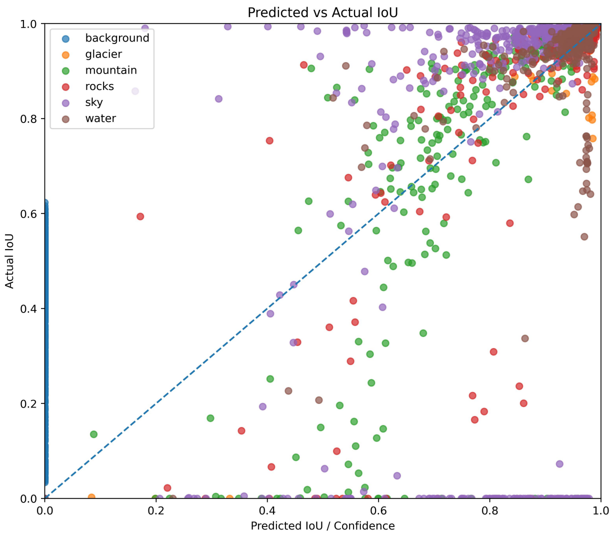
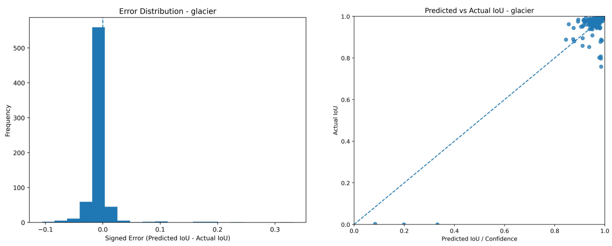


Figure 9. Predicted IoU versus actual IoU for all classes. The dashed diagonal represents perfect calibration. Clustering near the diagonal indicates strong agreement between predicted confidence and segmentation quality.



(a) Error distribution (b) Predicted vs. actual IoU

Figure 10. Confidence estimation analysis for the glacier class. (a) Error distribution showing a zero centered and sharply peaked distribution. (b) Predicted versus actual IoU, with strong clustering near the upper range indicating high segmentation accuracy and reliable confidence estimation.

mentation results with corresponding confidence estimates are provided in Fig. 11. These examples illustrate the model's behavior under varying environmental conditions, highlighting both stable predictions under moderate degradation and failure cases under severe visibility loss. The associated confidence scores further demonstrate the relationship between predicted segmentation quality and environmental difficulty.

5.4 Sensitivity Analysis of Model Configuration

All experiments in this section were conducted on a workstation equipped with an AMD Ryzen Threadripper 3990X (64 cores @ 2.9 GHz), 256 GB RAM, and two NVIDIA RTX A6000 GPUs (48 GB each).

5.4.1 Effect of Input Resolution To analyze the impact of input resolution on segmentation performance and computational efficiency, experiments were conducted using input sizes of 512×512, 1024×1024, and 1600×1600, using the SAM2 tiny variant.

Across all resolutions, the model maintains consistently high segmentation performance, with macro IoU values remaining within a narrow range of 0.95 ± 0.01, as summarized in Table 2. This indicates that pretrained SAM2 encoder features generalize effectively across varying spatial resolutions, with negligible impact on segmentation accuracy. For applications where scene structure remains relatively stable, lower input resolutions can therefore be used without loss of performance, enabling more efficient computation. In practice, reducing the input size to 512×512 results in approximately $2\times$ faster inference and significantly lower GPU memory usage, while higher resolutions increase computational cost accordingly.

Resolution	mIoU	Inf. (ms)	GPU Mem.
512	0.953	15.3	Low
1024	0.955	30.0	Medium
1600	0.956	69.7	High

Table 2. Segmentation performance and computational cost across input resolutions (SAM2 tiny variant).

5.4.2 Effect of Model Variants Experiments across SAM2 variants (tiny, small, base+, and large) show that segmentation performance remains comparable despite differences in model scale, as reported in Table 3. This behavior is attributed to the frozen image encoder, whose pretrained representations appear sufficiently expressive across variants, while adaptation is confined to the prompt encoder and mask decoder. As a result, increasing backbone capacity does not translate into significant task-specific gains, indicating that lightweight variants can achieve similar performance for glacier segmentation under this adaptation setup, supporting their use in resource constrained deployment scenarios.

Model	mIoU	Inf. (ms)	Train (min)
Tiny	0.952	30.0	13.7
Small	0.950	29.6	13.7
Base+	0.955	49.6	16.8
Large	0.956	67.2	18.3

Table 3. Comparison of SAM2 model variants under identical fine-tuning settings.

Adaptation	mIoU
Output layers only	0.74
+ Prompt (last)	0.82
+ Decoder attn	0.93
Full tuning	$\sim$0.95

Table 4. Layer-wise adaptation study of SAM2 components.

5.4.3 Layer-wise Adaptation Study To further understand the adaptation requirements of SAM2, a targeted ablation study was conducted by selectively fine-tuning different components of the mask decoder and prompt encoder, as shown in Table 4.

When fine-tuning was restricted to the output hypernetwork layers and the IoU prediction head, the model exhibited a significant drop in performance (mIoU = 0.74), indicating that these layers alone are insufficient to adapt the pretrained representations to the target domain. Extending the adaptation to include the final convolutional layer of the prompt encoder did not yield noticeable improvements, suggesting that prompt refinement alone is not the primary limiting factor.

In contrast, incorporating fine-tuning of the final attention layer within the decoder led to a substantial recovery in performance (mIoU = 0.93), approaching the results obtained with full decoder adaptation. This highlights the importance of attention

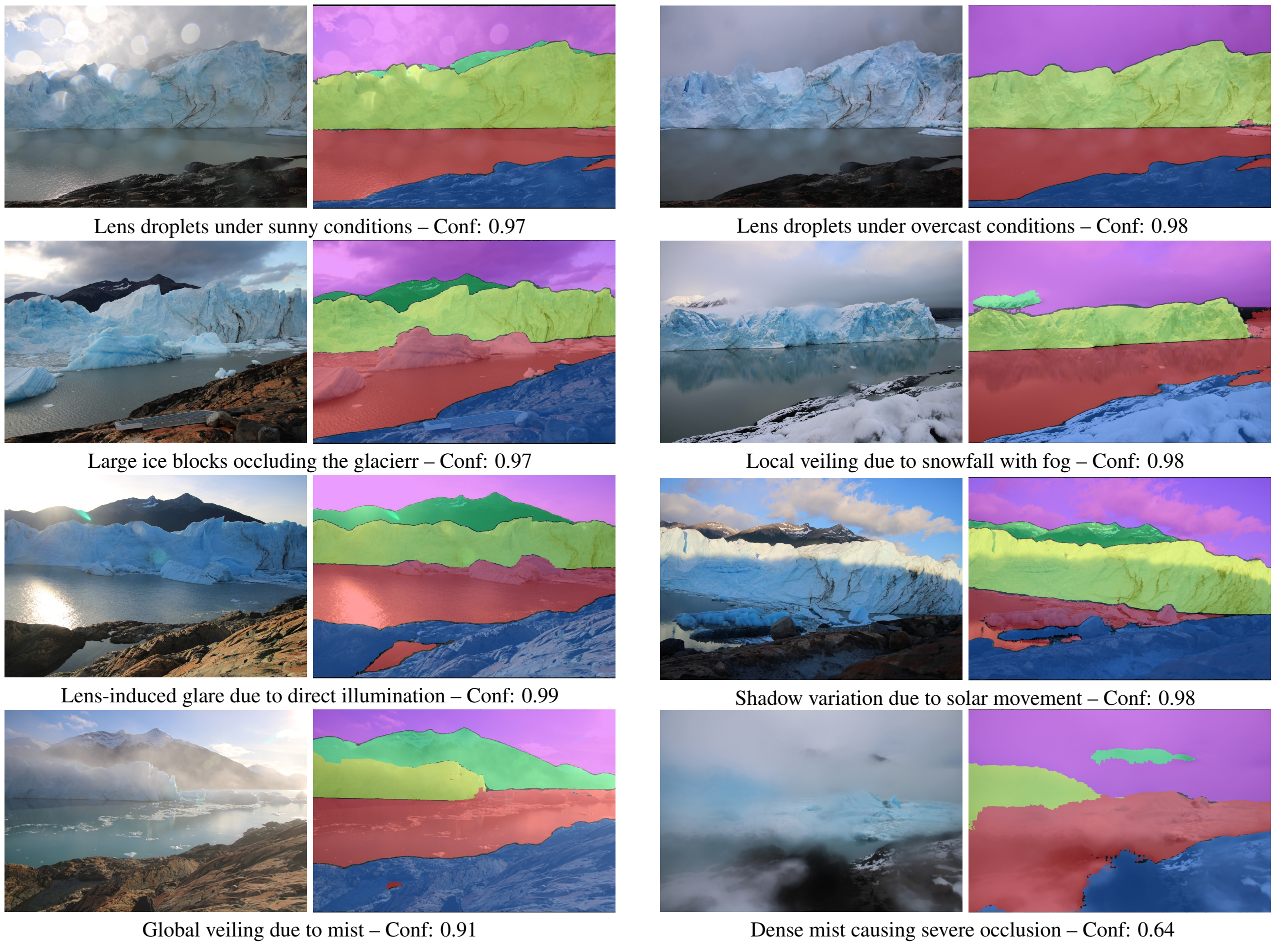


Figure 11. Qualitative segmentation results with predicted confidence scores only for the glacier class. Each pair shows the input image and the corresponding segmentation overlay. The confidence score reflects the model’s estimated segmentation quality of glacier. While high confidence aligns with stable predictions under clear and moderately degraded conditions, lower confidence is observed under severe visibility loss, corresponding to degraded segmentation performance.

level feature interaction in bridging the gap between pretrained representations and task specific segmentation requirements.

6. Discussion

This study does not aim to benchmark SAM2 against task-specific architectures, but rather to evaluate whether a general purpose foundation model can achieve sufficient performance under challenging environmental conditions. Within this scope, the focus is placed on understanding the model’s behavior under varying environmental and operational settings, rather than optimizing absolute performance against specialized baselines.

The experimental results demonstrate that the adapted SAM2 model provides stable and high quality segmentation under clear and moderately degraded conditions, across varying input resolutions and model configurations. Under severe visibility loss, performance degrades noticeably, which can be attributed to reduced feature distinguishability rather than model instability. The analysis of the confidence head shows that the model is generally well calibrated in estimating segmentation quality, with predicted IoU values closely aligned with actual IoU for most classes. In particular, for the glacier class, the error distribution is tightly centered and exhibits low variance, indicating reliable confidence estimation. However, certain failure modes remain, especially under challenging visual conditions, where high confidence may occasionally be assigned to lower quality predictions.

The analysis across resolution and model variants indicates that SAM2’s adaptation behavior is primarily governed by the pretrained encoder representations rather than model scale or input resolution. This suggests that, for visually consistent monitoring scenarios, efficient configurations can be adopted without compromising performance, supporting practical deployment under computational constraints.

The layer-wise ablation further highlights that effective adaptation does not require full model retraining, but instead depends on selectively updating components that mediate feature interaction. Overall, these observations demonstrate that SAM2 can be adapted in a controlled and efficient manner, where performance is maintained through targeted fine-tuning rather than increased model complexity.

7. Conclusion and Outlook

The presented adaptation of SAM2, combined with prompt generation and confidence estimation, demonstrates the feasibility of integrating such models into long term automated environmental monitoring pipelines. This configuration enables consistent mask generation while providing an internal mechanism to assess prediction reliability, which is essential for downstream processing. The results further indicate that such integration can be achieved without reliance on large model variants or extensive retraining, supporting efficient deployment in practical monitoring scenarios.

As a potential extension, the integration of additional lightweight modules, such as a classification head, could be explored to identify frames with insufficient visual quality prior to segmentation, reducing reliance on heuristic filtering methods. Furthermore, persistent deviations in predicted confidence may serve as an indicator of distribution shifts over time, potentially enabling adaptive model updates through selective fine-tuning on newly acquired data.

8. Acknowledgements and Declaration

This research work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – SFB/TRR 280; project ID: 417002380) and through the project “AI4Glaciers” (project no. 536233846). The authors acknowledge the use of ChatGPT and Codex-based tools for language refinement and programming assistance during manuscript preparation. All content has been reviewed and verified by the authors and the authors take full responsibility for the accuracy, integrity, and final content of the published work.

References

Bhattiprolu, S., 2024. Digitalsreeni image annotator.

Blanch, X., Eltner, A., Guinau, M., Abellan, A., 2021. Multi-Epoch and Multi-Imagery (MEMI) Photogrammetric Workflow for Enhanced Change Detection Using Time-Lapse Cameras. *Remote Sensing*, 13(8). https://www.mdpi.com/2072-4292/13/8/1460.

Blanch, X., Grundmann, J., Hedel, R., Eltner, A., 2026. AI image-based method for a robust automatic real-time water level monitoring: a long-term application case. *Hydrology and Earth System Sciences*, 30(3), 797–824. https://hess.copernicus.org/articles/30/797/2026/.

Duran Vergara, L. C., Blanch, X., Lokesh, B. N., Eltner, A., 2026. Challenges in automated 4D Point Cloud Generation for Glacier Calving Monitoring at high temporal Resolution.

Eltner, A., Kaiser, A., Castillo, C., Rock, G., Neugirg, F., Abellan, A., 2016. Image-based surface reconstruction in geomorphometry - merits, limits and developments. *Earth Surface Dynamics*, 4, 359-389.

Filho, L. C. C., Samatelo, J. L. A., Vassallo, R. F., 2025. Hierarchical vision transformer using shifted windows for skeleton-based action recognition. *2025 38th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 1–6.

Goli, L., Sabour, S., Matthews, M., Brubaker, M., Lagun, D., Jacobson, A., Fleet, D., Saxena, S., Tagliasacchi, A., 2024. Romo: Robust motion segmentation improves structure from motion.

Grothum, O., Epple, L., Bienert, A., Blanch Gorriz, X., Eltner, A., 2025. Near-continuous observation of soil surface changes at single slopes with high spatial resolution via an automated SfM photogrammetric mapping approach. *SOIL*, 11, 1007-1028.

James, M., Robson, S., Smith, M., 2017. 3-D uncertainty-based topographic change detection with structure-from-motion photogrammetry: precision maps for ground control and directly georeferenced surveys. *Earth Surface Processes and Landforms*, 42, 1769–1788.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything.

Lian, S., Li, H., 2024. Evaluation of segment anything model 2: The role of sam2 in the underwater environment.

Lokesh, B. N., Maas, H.-G., Eltner, A., 2026. A Scalable and Adaptable Framework for Domain-Specific Deep Learning Models with Improved Generalizability.

Mallalieu, J., Carrivick, J., Quincey, D., Smith, M., James, W., 2017. An integrated Structure-from-Motion and time-lapse technique for quantifying ice-margin dynamics. *Journal of Glaciology*, 63, 937-949.

Meena, T., Vijaya, J., Harsha, B., 2025. Swin transformers for remote sensing sar image classification. 1–6.

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., 2024. Sam 2: Segment anything in images and videos.

Schwalbe, E., Maas, H.-G., 2017. The determination of high-resolution spatio-temporal glacier motion fields from time-lapse sequences. *Earth Surface Dynamics*, 5, 861-879.

Shankar, S., Veen, C., 2023. Semantic segmentation of glaciological features across multiple remote sensing platforms with the Segment Anything Model (SAM). *Journal of Glaciology*, 70, 1-10.

Ulm, M., Elias, M., Eltner, A., Lotsari, E., Anders, K., 2025. Automated change detection in photogrammetric 4D point clouds – transferability and extension of 4D objects-by-change for monitoring riverbank dynamics using low-cost cameras. *Applied Geomatics*, 17.

Wagner, F., Eltner, A., Maas, H.-G., 2023. River water segmentation in surveillance camera images: A comparative study of offline and online augmentation using 32 CNNs. *International Journal of Applied Earth Observation and Geoinformation*, 119.

Xiang, Y., Zhao, L., Li, J., Gao, F., Bian, S., Hou, M., Luo, X., Guo, C., 2025. SAU-Net: A Deep Learning Approach for Glacier Mapping Based on Multisource Remote Sensing Data. *IEEE Access*, PP, 1-1.

Xie, B., Tang, H., Duan, B., Cai, D., Yan, Y., Agam, G., 2025. Masksam: Towards auto-prompt sam with mask classification for volumetric medical image segmentation.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8-36.

Zhu, Y., Locke, W., Yuan, J., Zhang, Y., Ma, Q., Liang, L., 2025. Leveraging SAM 2 and LiDAR for Automated Individual Tree Crown Delineation: A Comparative Evaluation of Prompting Methods. *Information Geography*, 1, 100025.