Response to comments by reviewer 2 (Pauline Delorme)

Reviewer comments are indicated in *blue and italic font*, while our responses are formatted as regular and black text. Our responses follow the reviewer comment's order that is structured in major and minor comments.

The manuscript presents ImageGrains 2.0, an updated workflow for automated grain segmentation in geoscientific imagery based on a fine tuned Cellpose SAM deep learning architecture. The authors adapt a biomedical segmentation model to detect sediment grains across a wide range of environments, including fluvial gravel, conglomerates, pro glacial deposits, and XR CT scans. The model is trained on an expanded and diversified dataset (IG2). Benchmarking against previous approaches demonstrates that the new model substantially improves segmentation accuracy and the reliability of grain morphometric measurements. The manuscript is generally well written and supported by clear illustrations. The proposed method represents a meaningful advance for automated grain size and shape analysis.
[…] This manuscript addresses an important methodological bottleneck in quantitative geomorphology. The dataset is rich, the benchmarking is solid, and the proposed workflow is a real improvement. However, the paper would benefit from clearer explanations of the model architecture, a more transparent comparison with previous work, and a deeper discussion of limitations, especially regarding small grains, dataset imbalance, and the accessibility of 3D methods.

We thank the reviewer for their astute analysis and for the very encouraging overall evaluation of our work.

Major comments
The introduction would benefit from a clearer explanation of why grain size and grain shape data matter for geomorphology, sediment transport, and landscape evolution. The authors cite relevant literature but do not articulate the broader implications (e.g., hydraulic roughness, abrasion, sediment sorting, transport thresholds, paleo environmental reconstruction). […]
We expand the introduction to include a broader view on the importance of accurate and reliable grain-size data in a geoscientific context (see also the responses to related comments from reviewer 1). We pick up this thread again in the discussion (see also related comments below).

I do not fully understand what the main improvement is compared to Mair et al. (2024). Figures 2 in Mair et al. (2024) and in this paper are very similar, and the text only briefly mentions the ViT backbone. I would appreciate if the authors provided more detail on the differences between Cellpose and Cellpose SAM.
ImageGrains 2.0 introduces two major improvements: an expanded dataset (IG2; 3 times the number of images in Mair et al., 2024; see also the response to the next comment below) and the new backbone architecture, Cellpose-SAM (Pachitariu et al., 2025), instead of Cellpose (Stringer et al., 2021). Cellpose-SAM uses an entirely different backbone segmentation

architecture (i.e., the modified transformer encoder of SAM instead of the much shallower U-net architecture of Cellpose). We add a brief statement in the introduction, and more details in the method section to make this clearer.

One of the major novelties of the paper appears to be the enriched dataset (IG2). In line 129, the authors claim that the dataset includes "the occurrence of vegetation…", however Figure 1 does not show any vegetation in the selected images. It would be useful to know what density or type of vegetation was included in the training data to better assess the model's ability to detect grains in vegetated channels.

We add a figure in the supplement that shows different vegetation and non-grain objects to illustrate these features.

For non specialists in deep learning methods, the paper is somewhat difficult to read. Section 2.3 (Cellpose SAM: re training and inference) is highly technical and hard to follow. The authors describe several modifications to the initial model but do not explain the relevance or motivation behind these changes.

We emphasize here that we use Cellpose-SAM (Pachitariu et al., 2025) without having customized the model architecture. Therefore, we only summarize the key features (and modifications compared to the original SAM architecture; Kirilov et al., 2023) of Cellpose-SAM to avoid even more technicalities. However, we add additional information and references throughout section 2.3 to highlight the effect of and motivation for these choices.

In Section 2.5 (Evaluating segmentation performance), lines 275–280, the authors state that they exclude grains smaller than 8 pixels. This seems reasonable to avoid misidentification, but in Figure 1 (subset NZ1), it appears that all the small particles composing the river bed are not detected and therefore not included in the statistics. This could be problematic for users who wish to apply the model output to physical modelling of river morphology.

There seems to be a misunderstanding. In Fig. 1, we show manually labelled grains for image tiles examples. Furthermore, NZ1 consists of images taken from near-vertical outcrops of lithified conglomerates (line 125), similar to images from subset FH_2. For these specific images, the resolution is not high enough to resolve grains of background sediment (mostly sand and very fine gravel). We have updated the figure caption and title to clarify that Fig. 1 shows labels (not predictions). Furthermore, we now emphasize the truncated nature of grain size distributions from images both in the method and discussion sections.

Figure 6 shows a comparison of many metrics, but some of them are not discussed in the main text (e.g., ΔIRn and Δeccentricity). The figure contains too many metrics that are not properly addressed, making it difficult to assess the improvement in size and shape accuracy.

We now expand the context and selection criteria for these metrics in section 2.2 (see the response to the related comment by reviewer 1). We also add titles to Fig. 6 and revise section 3.3 to name all metrics specifically.

Several figure captions are too short or incomplete. For example, Figure 2b is not described, and some figures lack information about what is being shown or how to interpret the metrics.
We carefully revise figure captions and add information where needed, and we add the missing caption for Fig. 2b.

Section 4.2.2 (3D Segmentation) is promising but underdeveloped. XR CT is expensive and not widely accessible. A comparison with more common 3D methods (SfM, point clouds, laser scanning) would strengthen the discussion. This limitation is acknowledged later (lines 546–550), but it should be integrated earlier and more explicitly.
Our main aim here is to introduce the 3D capacity of Cellpose-SAM for stacked image data, a data type obtained from XRD-CT scans. However, we consider reviewing and discussing different 3D methods beyond the scope of this study, and we do not specifically argue for (or against) using XRD-CT scanning. Instead, we focus only on summarizing key features and limitations of these data, and of alternative 3D data formats for grain segmentation. We revise the section to clarify this point early on.

Minor comments
Table S1 shows a strong imbalance between subsets (e.g., APF_2 has 59 tiles, CT only 6). This may bias the model toward fluvial gravel. The authors should discuss how this affects generalization.
We thank the reviewer for addressing the balance across subsets. While we have already discussed the intra-subset (in)balance (e.g., in sections 2.1, 4.1), we have not explicitly discussed the across-subset balance. The strong performance of our default Cellpose-SAM model across all subsets (cf. former Table 1, now Table 2), particularly in subsets with few images of different types (e.g., CT, DV_4), suggests that the inter-subset balance has little effect on Cellpose-SAM's generalization ability. We hypothesize that the ability of the SAM architecture to learn a multitude of representation categories, even with class imbalance (Kirilov et al., 2023), is the underlying cause for this performance. Therefore, the deep architecture, combined with the very general initial training of SAM and the 2 rounds of fine-tuning (first by Pachitariu et al. 2025 and then in our workflow; cf. Fig. 1), should effectively eliminate the need for inter-subset balance during model training. This is very much different for other, shallower architectures, such as the one used in Mair et al. (2024). One implication is that, for fine-tuning models to new image types, users would only need to annotate enough images for the model to work well on the new image type, and they would not need to annotate additional images to avoid across-subset imbalance. This could significantly reduce the number of required labels. We add statements to sections 3.2 and 4.2 to address this point.

Line 218: "if we left these out these data" → remove the extra "these".
Corrected.

Regarding the generalization capability (lines 319–320), the authors justify excluding S1_2 and PR for generalization testing, but the rationale is unclear.

We add a brief rationale statement. For the reasons for selecting these 2 subsets, we refer to lines 237f in section 2.4.

Lines 349–350: "Most notably, mean differences (including the ± one sigma standard deviation range) in grain size are below 12% for both the a and b axes…" This is incorrect when looking at Figure 6 (Δ mean b axis).
Corrected to: "mean differences in grain size are around 3% or below (with ± < 12% standard deviation for the one sigma range) for both the a- and b-axes".

Line 356: "For our default model, mean delta values are consistently below 2%…" This is incorrect for mean Δeccentricity and mean Δazimuth.
The mean delta values for our default model Cellpose-SAM (IG2) are -1.7 ± 2.1% for IR_n, -0.9 ± 0.9% for convexity, -1.5 ± 2.8% for eccentricity, and -1.4 ± 9.0 % for the azimuth (cf. Fig. 6). We add the specific numbers to communicate the variable standard deviation across metrics and to avoid confusion which number we are referring to here.

Figure S2: it would be helpful to add a ±10% shaded area for clarity.
Added (where applicable).

A discussion of the environmental impact of such methods would have been appreciated.
We thank the reviewer for raising this important point and we add a corresponding statement (for details, see the response to the related comment by reviewer 1).

Lines 585–587: "in a broad range of image datasets and applications" — please elaborate. Which types of applications? How might the limitations of the method affect these applications?
We intended to state here that our workflow could be adapted to segmentation tasks beyond the segmentation grains in clastic sediments, where segmentation performance remains a major bottleneck. We rephrase the statement and add some examples (including literature references).

Finally, as highlighted in the "future directions" section, one of the critical limitations of this method is the need for the existence or creation of a large dataset of manually annotated images (a substantial amount of work was required to create the IG2 dataset). This represents a considerable effort, and in many research contexts it is difficult to envision such extensive manual annotation as a feasible or scalable requirement.
First, all supervised deep learning methods are limited by the cost of producing ground truth data. Commonly, weakly-supervised or unsupervised methods are developed to circumvent this limitation. However, to test those, high-quality datasets are needed as benchmarks. Therefore, there should be an incentive to create high-quality data that is useful beyond a single application, which motivated us to make the full dataset available.
Second, our default model should generalize well across many settings, and if fine-tuning is required, it would likely require only a handful of annotated images (e.g., Mair et al., 2024;

Pachitariu and Stringer, 2022), or even fewer for the Cellpose-SAM architecture. Hence, the current tools are likely capable of achieving good results for individual research projects. Our argument is tailored towards a more general perspective, i.e., to improve segmentation models further, larger datasets are likely the area of most progress. These data would likely consist of a collection of smaller datasets from different studies, if published by the community, rather than of a single large labelling effort.
We adapt the corresponding section to more clearly reflect these notions.

References

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R.: Segment Anything, https://arxiv.org/abs/2304.02643, 2023.

Mair, D., Witz, G., Do Prado, A. H., Garefalakis, P., and Schlunegger, F.: Automated detecting, segmenting and measuring of grains in images of fluvial sediments: The potential for large and precise data from specialist deep learning models and transfer learning, Earth Surf. Process. Landf., 49, 1099–1116, https://doi.org/10.1002/esp.5755, 2024.

Pachitariu, M. and Stringer, C.: Cellpose 2.0: how to train your own model, Nat. Methods, 19, 1634–1641, https://doi.org/10.1038/s41592-022-01663-4, 2022.

Pachitariu, M., Rariden, M., and Stringer, C.: Cellpose-SAM: superhuman generalization for cellular segmentation, https://doi.org/10.1101/2025.04.28.651001, 1 May 2025.

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation, Nat. Methods, 18, 100–106, https://doi.org/10.1038/s41592-020-01018-x, 2021.