Efficient LoRA Adaptation of SAM3 for Interactive Bone Segmentation in Visible Human Cryosection Images

Philipi Schneider[1], **Carlos Eduardo Weiss**[1], **Eloiza de Oliveira Rosa**[1]
Vinícius Souza Capistrano[1], **Eliane Pozzebon**[1]

[1]Universidade Federal de Santa Catarina (UFSC)
Santa Catarina – Brazil

p.schneider@ufsc.br carlos.eduardo.weiss@grad.ufsc.br
eloiza.oliveira.rosa@grad.ufsc.br vinicius.capistrano@grad.ufsc.br
eliane.pozzebon@ufsc.br

Abstract. *This study evaluates Low-Rank Adaptation (LoRA) of SAM3 for interactive bone segmentation in Visible Human cryosection images. Adapters trained on Visible Korean bone data were externally evaluated on right-knee and right-foot volumes using idealized reference-derived point prompts and axial propagation. The best adapters increased volumetric Dice from 0.739 to 0.971 in the knee and from 0.411 to 0.917 in the foot while reducing severe failures and 3D fragmentation. These results indicate that parameter-efficient adaptation is effective under this idealized protocol, although performance varies with anatomy, training data, and adapter configuration.*

1. Introduction

Cadaver cryosectioning, initiated by the Visible Human Project (VHP) and continued in other projects such as the Visible Korean (VK), yields images capable of generating three-dimensional anatomical visualizations of excellent quality. However, doing so requires segmenting the anatomical structures in those images. The VHP, developed by the National Library of Medicine of the United States, produced reference anatomical volumes from cryosection, CT, and MRI images [Ackerman 1998, Spitzer et al. 1996]. The male specimen, used in the evaluation of this study, contains 1,878 axial color sections with 1.0 mm spacing. The VK, developed at Ajou University, provides high-resolution cryosection images with associated anatomical segmentations [Park et al. 2006, Park et al. 2005]. In this work, VK data from male human, female human, and canine specimens were used to train bone adapters, excluding from training the segmentations used in the external evaluation.

Despite the richness of these resources, complete volumetric segmentation of the VHP remains an open problem. Existing initiatives cover only specific regions or structures, such as the lower extremities provided by the University of Denver [Andreassen et al. 2023]. Producing these masks manually at scale is costly and requires specialized anatomical knowledge. Foundational segmentation models offer a promising basis for this interactive workflow. The Segment Anything Model (SAM) demonstrated that simple spatial prompts, such as points and boxes, can produce masks without category-specific training [Kirillov et al. 2023]. SAM 2 extended this approach to video, incorporating temporal memory to propagate masks along sequences [Ravi et al. 2024]. SAM3, used in this work, combines visual and conceptual prompts, a DETR-based image detector, and dense memory-based video tracking [Carion et al. 2025].

This design is well-suited to cryosection stacks: clicks on an anatomical structure can produce a mask per slice, and the axial sequence can be treated as video for propagation. However, foundational models trained predominantly on natural images show variable performance on medical images and frequently require domain adaptation [Mazurowski et al. 2023, Wu et al. 2023]. Anatomical cryosection images exhibit color, texture, scale, and spatial organization that differ from natural images, which justifies evaluating bone specialization.

A practical strategy for adapting large models is to use parameter-efficient fine-tuning techniques. Low-Rank Adaptation (LoRA) freezes the original model weights and learns low-rank updates in selected linear layers [Hu et al. 2022]. This reduces the number of trainable and stored task-specific parameters compared with full model fine-tuning, enabling the systematic testing of multiple experimental hypotheses: which training data to use, which modules to adapt, which rank to use, and at which depths of the visual backbone to apply adapters. Runtime or inference-speed gains were not evaluated.

This study investigates parameter-efficient specialization of SAM3 for interactive bone segmentation by training LoRA adapters on VK cryosection images and evaluating them on independently produced University of Denver labelmaps derived from the VHP. Its main contribution is a controlled comparison of training-data composition, adapter scope, rank, backbone depth, and axial propagation under this external VK-to-VHP protocol.

2. Data

This section describes the VK data used to train the adapters and the independent VHP-derived volumes used for external evaluation.

2.1. Training data

The adapters were trained on VK project data obtained from visualization software provided by Ajou University School of Medicine (`https://sites.google.com/ajou.ac.kr/anatomy`). Since the high-resolution raw data are not public, the images and segmentations included in those applications were used. The processed packages contained 1702, 1643, and 711 image-segmentation pairs for VK male, VK female, and VK dog, respectively. The masks were extracted by comparing pixels from the segmented images against the anatomical color catalog included in each package.

For this article, individual structures were grouped into a binary bone class comprising bones proper as well as cartilages, ligaments, and teeth. After extracting this composite class, slices without bone pixels were excluded, yielding 1690, 1635, and 692 eligible image-mask pairs for VK male, VK female, and VK dog. From these totals, 1352, 1308, and 554 pairs were used for weight updates, and 338, 327, and 138 were reserved for internal monitoring via a spatially distributed split along the axial sequence.

Color augmentation preserved the same masks and altered only the anatomical images. One augmented copy was generated per image, with a deterministic seed per source and slice. The transformations included independent variations of brightness and contrast at 1.0 ± 0.25, saturation at 1.0 ± 0.20, per-channel RGB gain at 1.0 ± 0.15, and gamma at 1.0 ± 0.20, preserving geometry, resolution, and image-to-mask alignment. The

augmented condition analyzed in this article combined the original male and female sets with their respective augmented copies, effectively doubling the training set size.

Training prompts were generated synthetically from the masks themselves. Positive points were sampled inside the mask and negative points in the near or far background from the structure. Each sample could contain up to five positive and three negative points, for a maximum of eight points, in accordance with comments in the official SAM3 implementation regarding training of the prompt encoder [Carion et al. 2025, Meta AI 2026]. The training set contained no human prompts nor propagated frames; therefore, the protocol represents interactive segmentation with idealized prompts derived from the ground truth.

2.2. External evaluation data: University of Denver / Visible Human

The main evaluation was performed on the public dataset from the University of Denver Center for Orthopaedic Biomechanics, "3D Models of the Visible Human Male and Female Lower Extremities" [University of Denver Center for Orthopaedic Biomechanics 2022], described by Andreassen et al. [Andreassen et al. 2023]. This dataset derives from cryosection images of the Visible Human Male and Female from the National Library of Medicine VHP [Ackerman 1995].

In this study we used the male specimen and the original University of Denver labelmaps as an independent external reference. Since these labelmaps were aligned to grayscale versions of the cryosections, they were re-aligned to the corresponding color photographs of the VHP male specimen, obtained from the DICOM collection of the NLM VHP in the NCI Imaging Data Commons [Clunie et al. 2024]. The alignment exploited the known correspondence between University of Denver frames and NLM slices, with localization by normalized cross-correlation after the necessary geometric transformations.

After alignment, the color images underwent white balance correction based on the 20-tone Kodak gray scale visible in the photographs. Per-channel RGB remapping curves were estimated from the neutral gray scale levels, reducing the yellow-greenish cast of the original images while preserving approximate luminance. The processed volume contains 1309 aligned axial slices, with axial spacing treated as 1 mm. The bone mask used in the evaluation was obtained by merging the bone labels of the right lower extremity.

3. Anatomical volume extraction

Anatomical volumes were extracted from the aligned University of Denver labelmaps by merging the relevant anatomical indices into a composite binary mask, followed by cropping using an aggregated bounding box with a fixed spatial margin. The right knee combined the right femur, tibia, fibula, and patella across 350 contiguous slices, frames 504–853 (NLM slices 1074–1423), with 1 mm axial spacing and a crop of 705×747 pixels. The right foot was extracted from the distal segment of the lower extremity; after excluding four terminal slices without a positive bone mask, the evaluation used 176 positive slices, frames 1129–1304 (NLM slices 1699–1874). The aligned color images and corresponding masks were exported slice by slice, preserving pixel-to-pixel correspondence between image and ground truth.

The knee and foot were prioritized as the anatomically most complex and challenging regions of the available lower extremity. The pelvis was excluded because the corresponding slices contain vertebral column bones without masks in the University of Denver dataset, which did not segment those structures, making them unsuitable for evaluation. The intermediate segments — thigh and leg — present bones with simpler contours and more continuous sequences, less representative of the boundary cases that adapters must handle.

4. Model and adapters

SAM3 combines a prompt-conditioned DETR-based image detector [Carion et al. 2020, Carion et al. 2025] with a memory-based video tracker. The detector produces image masks, while the tracker propagates masks and visual prompts through a sequence using reference frames stored in memory, following the approach introduced by SAM2 [Ravi et al. 2024, Carion et al. 2025].

The detector and tracker share a visual backbone. In this article, visual backbone denotes the part of the model that transforms the input image into feature maps used by downstream detection, segmentation, and propagation modules. LoRA adaptation was applied to the SAM3 detector, but restricted to visual backbone modules selected by the pattern `backbone.vision_backbone.trunk.blocks.*`. The tracker, the prompt encoder, and the mask decoder remained frozen. The prompt encoder converts input points into vector representations used by the model; the mask decoder combines these representations with backbone feature maps to produce the final mask.

LoRA was used as a parameter-efficient fine-tuning approach: the original SAM3 weights were kept frozen, and only low-rank matrices added to selected linear layers were trained [Hu et al. 2022]. The rank `r` controls the adapter capacity: larger values increase the number of trainable parameters; smaller values restrict the update, reducing training cost and storage. In this study, `r8/a16` indicates rank 8 with `lora_alpha = 16`, and `r16/a32` indicates rank 16 with `lora_alpha = 32`.

In the implementation and checkpoint used in this study, the visual backbone exposed for LoRA adaptation contains 32 Transformer blocks (blocks 0–31). Four operational subsets were defined: `early` (blocks 0–7, 8 blocks), `middle` (blocks 8–23, 16 blocks), `late` (blocks 24–31, 8 blocks), and `all` (all 32 blocks). Three target module scopes were evaluated: Attention (`attn.qkv` and `attn.proj`), MLP (`mlp.fc1` and `mlp.fc2`), and Attention+MLP (all four modules combined).

Here, `MLP` denotes the position-wise feed-forward submodule of each Transformer block. Its `mlp.fc1` layer expands the feature dimension and `mlp.fc2` projects it back, transforming features after attention has combined information across image positions.

In the attention experiments, adapters were inserted in the `attn.qkv` linear projections, which generate queries, keys, and values, and in the output projection `attn.proj`. In the MLP experiments, adapters were inserted in the `mlp.fc1` and `mlp.fc2` layers.

5. Experimental configurations

Ten bone adapter families were evaluated, in addition to the SAM3 baseline. For each family, the available epoch checkpoints were analyzed:

Main variable	Configuration
Baseline	SAM3 base
VK male only	Attention, all blocks, r8/a16
Early blocks	VK male only, Attention, blocks 0–7, r8/a16
Middle blocks	VK male only, Attention, blocks 8–23, r8/a16
Late blocks	VK male only, Attention, blocks 24–31, r8/a16
VK male + VK female	Attention, all blocks, r8/a16
Higher rank	VK male + VK female, Attention, all blocks, r16/a32
VK dog	VK male + VK female + VK dog, Attention, all blocks, r8/a16
MLP-only	VK male + VK female, MLP, all blocks, r8/a16
Attention+MLP	VK male + VK female, Attention+MLP, all blocks, r8/a16
Color augmentation	VK male + VK female (original) + VK male + VK female (augmented), Attention, all blocks, r8/a16

In the slice-by-slice test, each volume was evaluated with 121 candidates: the SAM3 baseline and 120 epoch checkpoints.

6. Training

Training simulated interactive segmentation. Each sample began with a positive point inside the ground truth mask; intermediate predictions were then used to add correction points in false negative and false positive regions. The protocol allowed up to eight points per sample, including the initial point and points added during up to four refinement rounds.

The loss function was the sum of soft Dice loss and binary cross-entropy with logits ($\mathcal{L}_{\text{bce}}$). The Dice loss favors global overlap between the predicted mask and the ground truth, while binary cross-entropy penalizes local per-pixel errors:

$$\text{Loss} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i + 10^{-6}} + \mathcal{L}_{\text{bce}}(z, g),$$

where z_i is the predicted logit for pixel i, $p_i = \sigma(z_i)$ is the probability after the sigmoid, and $g_i \in \{0, 1\}$ is the reference mask value.

Images were resized to 1008×1008 pixels before input to SAM3, the resolution used by the official model implementation. `lora_dropout = 0.05` and up to four refinement rounds were used, with early stopping when the intermediate prediction reached an IoU of 0.85 with respect to the ground truth mask. Final masks were interpreted at the original resolution of each slice.

All adapter families were trained for 12 epochs with AdamW, a constant learning rate of 2×10^{-5}, weight decay of 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. No learning-rate scheduler was used. Training used one image per step, bfloat16 automatic mixed precision with gradient scaling, and gradient clipping with maximum norm 1.0. The

SAM3 mask decoder, prompt encoder, tracker, and original backbone weights remained frozen; only the LoRA parameters were updated.

All adapter families used `gradient_accumulation_steps = 2`, corresponding to an effective batch size of two images, with gradient checkpointing disabled. Seed 20260609 was recorded for all runs except the rank-16 and canine-data configurations, whose metadata did not preserve the seed.

6.1. Computational implementation

The pipeline was implemented in Python with PyTorch and PEFT for training and inference of SAM3 with LoRA adapters; standard computer vision and numerical analysis libraries were used for preprocessing, evaluation, and figure generation. Training runs were executed on two machines — an NVIDIA DGX Spark (Linux, initial phase) and a computer with an NVIDIA GeForce RTX 5070 Ti GPU (Windows 11, 32 GB RAM) — with the same experimental configurations in both environments.

7. Evaluation protocols

Two complementary protocols measured independent prompted segmentation and memory-based propagation through the axial sequence.

7.1. Slice-by-slice segmentation

Each slice was evaluated independently with prompts derived from the ground truth mask — positive points sampled inside the mask and negative points in the nearby background, without human intervention. The protocol therefore represents an idealized upper bound for interactive use rather than an estimate of automatic segmentation or observed human performance. Because reference-derived points can be placed systematically inside the target and away from ambiguous boundaries, prompts supplied by real operators may produce lower performance. For each prediction, only the final binary masks were saved, without storing intermediate model outputs. The predicted masks were exported in the same format as the reference masks: grayscale PNG, with value 0 for background and 255 for pixels classified as bone.

7.2. Video propagation

For each adapter family, the epoch with the highest slice-by-slice volumetric Dice on the external evaluation volume was retrospectively selected for propagation. Because selection used the evaluation volume itself, these results characterize the best checkpoint observed retrospectively rather than a prospective selection on unseen data. The SAM3 baseline was also evaluated. Three propagation protocols were then run: anchor at the first slice and propagation to the end, anchor at the last slice and propagation to the beginning, and central anchor with propagation to both extremes. Thirty-three propagation sequences per volume were evaluated, corresponding to 11 candidates/families and three protocols.

8. Metrics

For each slice t, the predicted mask P_t was compared against the reference G_t, classifying each pixel as true positive (TP), false positive (FP), or false negative (FN). Dice was computed as:

$$\text{Dice} = \frac{2\,TP}{2\,TP + FP + FN}.$$

The `area_ratio` is the ratio between the predicted area and the reference area; values above 1 indicate oversegmentation, below 1 undersegmentation. Mean Dice is the arithmetic mean of per-slice Dice scores; it differs from volumetric Dice because it assigns equal weight to each slice regardless of area. The number of slices with Dice below 0.90 was also counted as an indicator of localized failures.

For the stacked volume, volumetric Dice aggregates TP, FP, and FN across all slices before computing the score, naturally weighting slices with larger bone areas:

$$\text{Volumetric Dice} = \frac{2\sum_t TP_t}{2\sum_t TP_t + \sum_t FP_t + \sum_t FN_t}.$$

The volumetric area ratio was computed analogously.

For spatial continuity, the axial area variation error (`p95_area_delta_error`) and the centroid displacement error between consecutive slices (`p95_centroid_step_error_px`) were summarized by the 95th percentile. For fragmentation, the stacked masks were evaluated by 3D connected components with 26-neighbor connectivity; the number of components indicates disconnected islands or broken masks in the volume.

9. Results

Results are organized by overall adaptation gain, training-data composition, LoRA configuration, checkpoint selection, propagation, and volumetric coherence.

9.1. Overall gain over the SAM3 baseline

The bone-specialized LoRA adapters substantially improved the SAM3 baseline on both evaluated volumes. In the knee, the SAM3 baseline achieved a volumetric Dice of 0.7393, with severe oversegmentation; the best adapter, with color augmentation (epoch 12), reached 0.9714 (+0.2321). In the foot, the baseline was even weaker (0.4114); VK male + VK female attention-only (epoch 5) and MLP-only (epoch 9) were effectively tied at 0.9168 and 0.9166, respectively (approximately +0.505 over the baseline).

Beyond the volumetric Dice gain, the adapters reduced the number of slices with Dice < 0.90 and corrected the SAM3 baseline oversegmentation, bringing the area ratio close to 1.0 in the best results.

Figure 1 qualitatively illustrates this contrast in a representative knee slice (frame 0849, NLM 1419). The SAM3 baseline shows severe oversegmentation, marked by the predominance of false positives in red. The color augmentation adapter (epoch 12) identifies bone regions with greater precision, concentrating residual errors at structural boundaries.

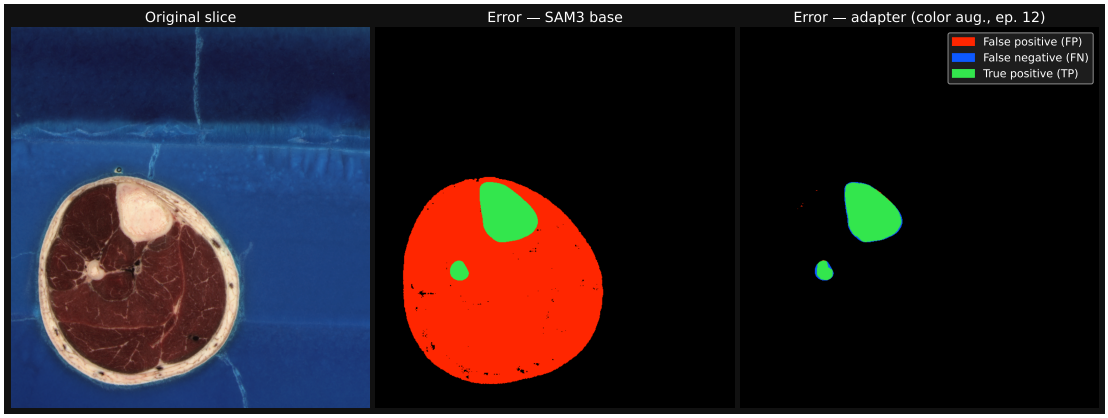


Figure 1. Segmentation errors in the right knee, frame 0849 (NLM 1419). Left to right: original slice, SAM3 baseline error mask, and error mask of the color augmentation adapter (epoch 12).

9.2. Effect of training dataset

The first experimental axis evaluated how training data composition affects external segmentation on the VHP. All configurations in this section use LoRA on attention, all visual backbone blocks, and rank 8/alpha 16; the main variable is the training set.

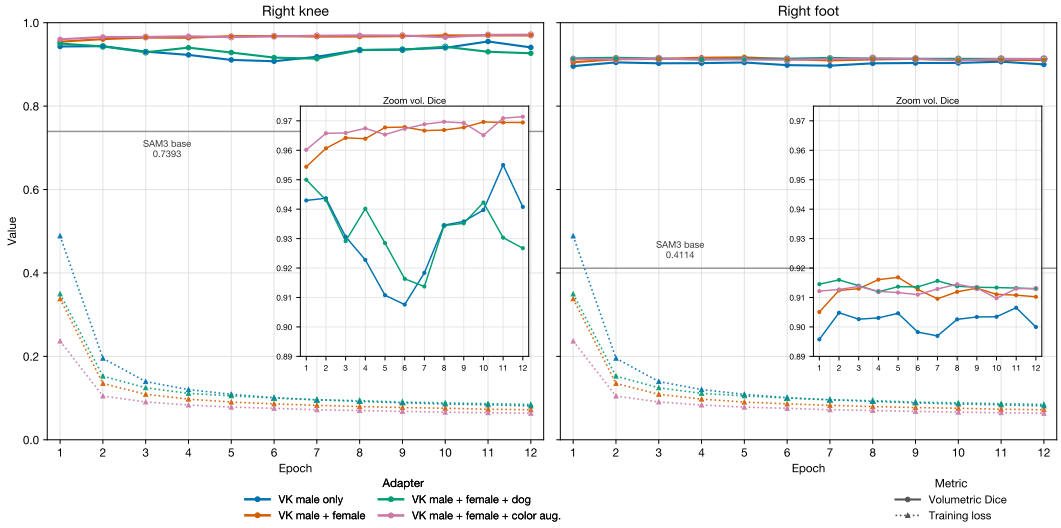


Figure 2. Training loss and volumetric Dice per epoch — dataset composition.

Figure 2 shows the evolution of both metrics per epoch. Training loss is shown once per adapter; volumetric Dice is displayed separately for knee and foot on the same axis, making clear that its decrease does not imply monotonic improvement in external Dice. The inclusion of the female specimen was the most consistently useful modification, particularly in the knee. Color augmentation produced the best result in that region but had little additional effect on the foot. The inclusion of VK dog data showed region-dependent behavior: it hurt the knee relative to the male+female human training but remained competitive in the foot. Taken together, these results indicate that additional data diversity is not automatically beneficial; its effect depends on the anatomical region and the similarity between the added structures and the evaluated volume.

9.3. Effect of LoRA configuration

The second experimental axis evaluated changes to the adapter itself: LoRA capacity, architectural scope, and depth of the adapted blocks.

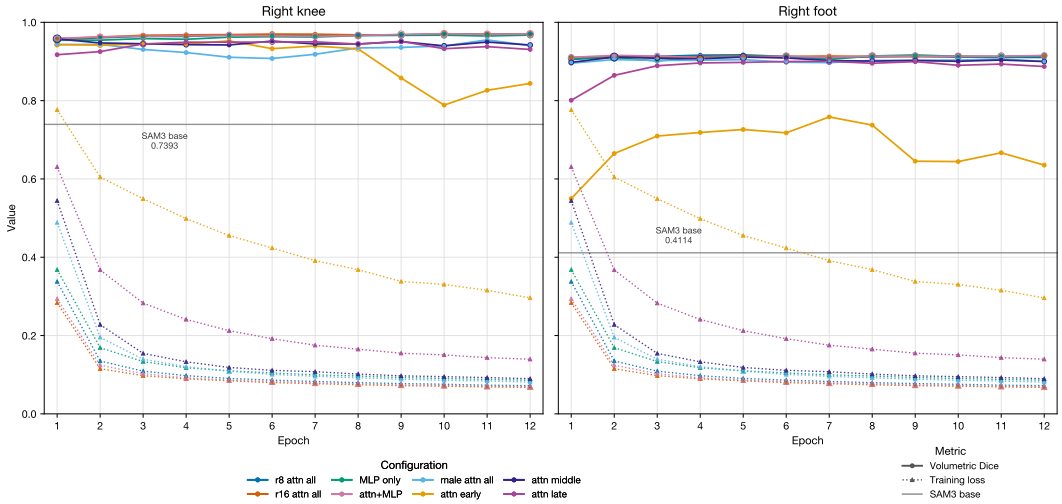


Figure 3. Training loss and volumetric Dice per epoch — LoRA configurations.

Figure 3 brings together the comparisons of rank, architectural scope, and block depth; the horizontal line marks the volumetric Dice of the SAM3 baseline in each volume. Increasing rank from r8/a16 to r16/a32 produced a small gain in the knee and no improvement in the foot, suggesting that additional capacity was not the main limiting factor in this battery. The contrast between attention-only, MLP-only, and Attention+MLP showed similar performance across all three scopes. The most relevant point is that MLP-only did not function as a weak control: it was competitive in the knee and was among the best results in the foot. This suggests that domain adaptation to the anatomical domain does not depend exclusively on attention layers.

The middle subset obtained the highest observed performance among the attention depth variants. However, it adapts 16 blocks, whereas the early and late subsets adapt only eight; consequently, comparisons involving the middle subset confound backbone location with trainable capacity. Only the early-versus-late comparison holds the number of adapted blocks constant.

9.4. Epoch selection and stability

Evaluating all epochs revealed that the final checkpoint is not always the best. In the knee, the best adapters were relatively stable up to epoch 12. In the foot, several configurations reached their best performance at intermediate epochs. In the knee, color augmentation and Attention+MLP achieved their best result at the final epoch. In the foot, VK male + VK female attention-only peaked at epoch 5, MLP-only at epoch 9, VK male + VK female + VK dog at epoch 2, and early attention at epoch 7, with subsequent degradation relative to the best checkpoint.

9.5. Video propagation

Video propagation was evaluated using the best epoch of each adapter family by the slice-by-slice volumetric Dice criterion. Three protocols were tested: propagation from the

first slice to the end, from the last slice to the beginning, and from the central slice to both extremes.

In the knee, the best propagation among adapters achieved volumetric Dice of 0.9648 with VK male + VK female attention-only (epoch 10) under central bidirectional propagation, close to its slice-by-slice result (0.9696). In the foot, the best propagated result was 0.9066, also close to the best slice-by-slice value from the corresponding family (0.9156), but with greater sensitivity to propagation direction. The SAM3 baseline improved substantially with propagation, rising from 0.7393 to 0.9363 in the knee and from 0.4114 to 0.8657 in the foot. These results suggest that temporal memory stabilizes segmentations prone to severe failures when evaluated slice by slice, although the best propagated adapters still outperform the propagated SAM3 baseline.

9.6. Spatial continuity and fragmentation

Continuity metrics show that volumetric accuracy and axial coherence are not identical. The foot showed greater axial area variation error, greater centroid displacement error, and greater fragmentation than the knee, consistent with its more discontinuous anatomy: the knee ground truth contained 3 3D connected components, while the foot contained 20.

Propagation strongly reduced 3D fragmentation in both volumes, although it may lose Dice relative to the best slice-by-slice segmentations. In the foot, this reduction should be interpreted together with Dice and anatomical inspection, as fewer components may also reflect inappropriate fusion between adjacent bones.

Figure 4 illustrates the volumetric impact of this difference in 3D renderings of the right knee, generated with a custom viewer. The SAM3 baseline produces a volume with evident fragmentation and axial discontinuity. The color augmentation adapter reproduces the anatomical structure with continuity close to the ground truth, reflecting the correction of oversegmentation documented by the quantitative metrics.



Figure 4. Volumetric renderings of the right knee, generated with a custom viewer.

10. Discussion

Bone-specialized LoRA adapters substantially improve SAM3 on VHP cryosection images, although the gain depends on region. In the knee, they correct baseline oversegmen-

tation and raise volumetric Dice to approximately 0.971. In the foot, where the baseline fails more broadly, the absolute gain is even larger.

The study also shows a clear dependence on anatomical region and metric. The color augmentation configuration achieved the highest observed volumetric Dice in the knee, while VK male + VK female attention-only, `MLP-only`, and `VK male + VK female + VK dog` produced effectively tied high foot scores. Moreover, adapters that lead in volumetric Dice are not necessarily those producing the best axial continuity or lowest fragmentation: video propagation improves fragmentation but may reduce Dice relative to slice-by-slice segmentation. Thus, one region is insufficient for ranking adapters, and data and target variations capture different properties: attention-only was strong in the knee but not clearly superior to MLP-only and Attention+MLP, while MLP and anatomical variation appear more relevant for the small, fragmented structures in the foot.

Axial propagation improves even the baseline and keeps the best adapters close to slice-by-slice performance, while markedly reducing connected components. However, it is sensitive to anchor and direction, especially in the foot. Continuity metrics therefore complement Dice: methods with similar overlap may differ in area variation, centroid jumps, holes, islands, and fragmentation.

11. Limitations

The experiments used two volumes from one male specimen and one run per configuration. Fine-grained differences are therefore observed results, not statistically established rankings. Seeds were not preserved for the rank-16 and canine-data runs, precluding seed-level variance estimates for those two families. Other lower-extremity regions were excluded because the pelvis reference omits some bones, whereas thigh and leg are anatomically simpler; the available reference may also reflect anatomy-specific labeling choices.

Reference-derived prompts and retrospective epoch selection on the evaluation volume make this an idealized upper-bound analysis rather than a prospective deployment protocol. Moreover, LoRA was compared with unadapted SAM3, but not partial or full fine-tuning or another medical adapter; the study establishes feasibility and parameter efficiency, not superiority over those alternatives. Future work should include repeated runs, human prompts, surface distances, additional specimens, anatomical groups and modalities, and possible degradation outside the bone domain at late epochs.

12. Conclusions

Bone-specialized LoRA adapters improved SAM3 segmentation on VHP cryosection images, reaching volumetric Dice of approximately 0.971 in the knee and 0.917 in the more challenging foot, although axial continuity was lower in the foot.

Within this setting, LoRA was a viable parameter-efficient strategy, although the highest observed configuration depended on region and metric. The highest score did not always occur at the final epoch, while prospective use requires selection independent of the test volume. Volumetric and continuity metrics complement one another in axial sequences. Propagation reduced fragmentation but must be evaluated separately from slice-by-slice segmentation.

References

Ackerman, M. J. (1995). Accessing the visible human project. *D-Lib Magazine*.

Ackerman, M. J. (1998). The Visible Human Project. *Proceedings of the IEEE*, 86(3):504–511.

Andreassen, T. E., Hume, D. R., Hamilton, L. D., Walker, K. E., Higinbotham, S. E., and Shelburne, K. B. (2023). Three dimensional lower extremity musculoskeletal geometry of the Visible Human female and male. *Scientific Data*, 10:34.

Carion, N. et al. (2025). SAM 3: Segment anything with concepts. arXiv:2511.16719.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229.

Clunie, D. et al. (2024). DICOM converted images for the NLM-Visible-Human-Project collection. Zenodo. `https://doi.org/10.5281/zenodo.12690050`.

Hu, E. J. et al. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Kirillov, A. et al. (2023). Segment Anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003.

Mazurowski, M. A. et al. (2023). Segment Anything Model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918.

Meta AI (2026). Segment Anything Model 3 official implementation. `https://github.com/facebookresearch/sam3`.

Park, J. S., Chung, M. S., Hwang, S. B., Lee, Y. S., Har, D.-H., and Park, H. S. (2005). Visible Korean Human: Improved serially sectioned images of the entire body. *IEEE Transactions on Medical Imaging*, 24(3):352–360.

Park, J. S., Chung, M. S., Hwang, S. B., Shin, B.-S., and Park, H. S. (2006). Visible Korean Human: Its techniques and applications. *Clinical Anatomy*, 19(3):216–224.

Ravi, N. et al. (2024). SAM 2: Segment anything in images and videos. arXiv:2408.00714.

Spitzer, V., Ackerman, M. J., Scherzinger, A. L., and Whitlock, D. (1996). The visible human male: A technical report. *Journal of the American Medical Informatics Association*, 3(2):118–130.

University of Denver Center for Orthopaedic Biomechanics (2022). 3D models of the Visible Human male and female lower extremities. Dataset. `https://doi.org/10.56902/COB.vh.2022.0`.

Wu, J. et al. (2023). Medical SAM Adapter: Adapting segment anything model for medical image segmentation. arXiv:2304.12620.