Adapting Segmentation Foundation Models to 3D Pulmonary Vasculature: A Translational Research Agenda

Literature review and conceptual analysis

Abstract

Segmentation foundation models promise reusable representations, broad task coverage, and reduced dependence on task-specific training, but pulmonary vasculature exposes a difficult domain gap. Chest computed tomography is volumetric, often anisotropic, and visually dominated by structures larger than the peripheral vessels of interest. Pulmonary vessels form a sparse branching network whose clinical value depends on connectivity, caliber, and anatomical identity rather than foreground overlap alone. This research agenda examines how foundation models can be adapted to this setting without discarding their pretrained knowledge or incurring the full cost of high-resolution three-dimensional attention. MorVess is used as a concrete case because it freezes a two-dimensional Segment Anything Model encoder, adds a lightweight 2.5D adapter, predicts mask, distance, and thickness fields, and applies global-local fusion with staged optimization. The design illustrates a productive compromise between general representation and specialized geometry, while its limited depth context and sensitivity to anisotropic voxels reveal open problems. The agenda proposes research priorities in volumetric context, parameter-efficient adaptation, anatomy-aware prompting, annotation design, cross-domain evaluation, uncertainty, and clinical workflow integration. It argues that successful translation requires a model to preserve vascular structure across scanners and diseases, produce inspectable uncertainty, and support downstream measurements or corrections. Foundation-model adaptation should therefore be evaluated as a complete evidence and workflow problem, not only as a competition for higher Dice.

The Promise and the Domain Gap

Foundation models are trained on broad data so that their representations can be reused across tasks. Segment Anything demonstrated this idea for promptable two-dimensional image segmentation, pairing an image encoder with prompt and mask decoders and training on a dataset exceeding one billion masks (Kirillov et al., 2023). Its interface allows points or boxes to specify an object, and its zero-shot behavior is impressive on many natural images. Medical use, however, introduces unfamiliar intensity statistics, weak boundaries, modality-specific artifacts, and semantic targets that may not resemble discrete photographed objects.

MedSAM responds to that domain gap through large-scale medical fine-tuning. Ma et al. (2024) assembled more than 1.5 million image-mask pairs across multiple modalities and evaluated internal and external tasks. The study shows that medical adaptation can improve performance and generalization relative to unmodified SAM. Yet its curation excluded some tiny or branching targets, underscoring a central problem: a universal medical segmentation representation is not automatically a pulmonary vessel representation. Small tubular structures require sensitivity to features that broad training may underweight.

Fully volumetric foundation modeling offers a complementary route. Wang et al. (2025) developed SAM-Med3D as a general-purpose model for three-dimensional medical segmentation, demonstrating that SAM-style representation learning can be reformulated around volumetric inputs. Such a model preserves depth context more directly than slice-wise adaptation, but processing large chest CT volumes still forces choices about patch size, resolution, and memory. Pulmonary vessels therefore provide a useful test of whether volumetric generality survives at the fine spatial scale of distal branches.

The classical U-Net remains relevant here. Ronneberger et al. (2015) designed skip-connected encoding and decoding to combine semantic context with localized detail. 3D U-Net extended the pattern to volumetric data and showed how sparse slice annotations could support dense prediction (Cicek et al., 2016). nnU-Net later demonstrated that preprocessing, spacing, patch size, network depth, and training configuration can be adapted systematically to each biomedical dataset (Isensee et al., 2021). These systems are task-specific, but they embody decades of useful inductive bias. The research question is not whether foundation models make them obsolete. It is how broad pretrained features can be combined with volumetric and morphological structure more effectively than either approach alone.

MorVess as an Adaptation Blueprint

MorVess offers a concrete blueprint for this combination. Mao et al. (2026) retain a frozen SAM vision transformer encoder and insert a lightweight 2.5D adapter that processes five neighboring CT slices. A convolution spanning the depth direction communicates local inter-slice information, and its output is added residually to the planar features. The decoder jointly predicts the vessel mask, a vessel distance map, and a vessel thickness map. A global-local fusion block uses multiple feature levels and geometric outputs to reconstruct fine branches that a low-resolution foundation-model decoder might otherwise miss.

The geometric targets are not generic auxiliary tasks. The distance map encodes proximity to the vessel boundary through a continuous field. The thickness map propagates centerline radius through the lumen and expresses the expectation that caliber changes coherently along a tree. Training combines cross-entropy, Dice, centerline Dice, distance regression, and scale-normalized thickness regression. A two-stage schedule first adapts macro features and inter-slice information, then freezes the adapter and refines fusion and geometric prediction at a different learning rate.

This structure addresses three constraints at once. Freezing the encoder protects pretrained knowledge and reduces the number of trainable parameters. The adapter restores some volumetric context without processing the full CT with three-dimensional self-attention. The geometric heads compensate for the generic encoder's lack of explicit knowledge about tubular continuity and caliber. Mao et al. (2026) report approximately one million trainable parameters, 42 GMACs per five-slice stack, and 4.2 GB peak memory under their stated comparison, substantially below the nnU-Net and diffusion-model baselines listed in the study.

The reported evidence is correspondingly broad. MorVess is compared on Parse2022 pulmonary arteries and AIIB2023 chest structures using Dice, clDice, HD95, missing-rate and branch-oriented metrics. It also evaluates total vessel volume, caliber distributions, tortuosity, and small-vessel fraction, and tests transfer without fine-tuning to HiPas pulmonary veins and ATM2022 airways. The design therefore serves as more than a model proposal: it is a hypothesis that general visual representations become useful for vessels when local volumetric context and differentiable morphology are supplied.

Priority 1: Expand Context Without Losing Resolution

The most immediate limitation is depth context. MorVess uses two slices before and after a target, so its direct Z-axis receptive field is short. Mao et al. (2026) explicitly note that this limits long-range dependencies spanning tens of millimeters and reduces robustness to anisotropic voxels. A vessel that curves steeply through the volume may leave the local stack, while thick slices make adjacent images farther apart physically than the architecture assumes.

Full volumetric models address this issue but consume substantial memory at thoracic resolution. Swin UNETR uses hierarchical shifted-window transformers for volumetric medical segmentation, combining local windows with cross-window communication (Hatamizadeh et al., 2022). Such architectures suggest one path: encode coarse three-dimensional context at lower resolution while retaining a high-resolution two-dimensional or sparse decoder for vessels. The model could use global lung coordinates, lobe masks, and central-vessel graphs to condition local predictions.

Another path is recurrent or state-space processing along slices, where a compact memory carries branch information farther than a fixed five-slice window. Sparse attention can operate on candidate vessel tokens rather than every voxel. Multi-plane processing can combine axial, coronal, and sagittal views, although inconsistent view fusion may create duplicated or broken branches. The agenda should compare these mechanisms under equal memory and latency budgets, because a more expensive model is not necessarily a more deployable one.

Physical spacing must enter the architecture. Depth offsets should represent millimeters, not only slice indices. Training can sample neighbors based on physical distance and inform the adapter of spacing. Anisotropy-aware kernels or resampling uncertainty may prevent a model from treating interpolated detail as observed anatomy. Evaluation should include native thick-slice scans rather than only isotropically resampled volumes.

Priority 2: Make Parameter Efficiency Anatomically Selective

Parameter-efficient adaptation restricts which pretrained weights can change. Low-rank adaptation was introduced as a way to update large models through small trainable matrices rather than full fine-tuning (Hu et al., 2022). Although the original application was language modeling, the principle has spread to vision encoders. For pulmonary vessels, the research question is where adaptation capacity yields the greatest anatomical value.

A uniform low-rank update may spend capacity on dominant textures or central vessels. Rank could instead vary by encoder depth, scale, or branch difficulty. Early layers may need limited adjustment to CT intensity and noise; intermediate layers may need stronger adaptation for tubular orientation; later layers may need semantic conditioning for artery, vein, airway, or lesion relationships. Sensitivity analysis can identify which blocks influence distal recall, connectivity, and caliber rather than only average Dice.

MorVess uses a rank-selection experiment and finds diminishing gains beyond a chosen setting, which is a useful start. Future studies should report performance per trainable parameter, peak memory, throughput per volume, and energy under standardized hardware. More importantly, they should report structural return on capacity: change in small-branch recall, clDice, and diameter error for each adaptation budget. A compact model is valuable only if it preserves the features required by the clinical task.

Adapters also offer modularity. Separate lightweight modules could specialize for contrast-enhanced CT, non-contrast CT, low-dose scans, fibrosis, or postoperative anatomy while sharing one backbone. The risk is fragmentation into poorly calibrated variants. A routing mechanism should select or combine adapters based on acquisition metadata and uncertainty, and external evaluation should test whether the selected module truly matches the case.

Priority 3: Use Pretraining That Respects Volumetric Structure

Natural-image pretraining supplies strong general features but little direct knowledge of CT physics or three-dimensional continuity. Masked autoencoders learn representations by reconstructing hidden image patches and can scale effectively with unlabeled data (He et al., 2022). A pulmonary foundation model could pretrain on large chest CT collections using masked volumetric or multi-slice reconstruction, spacing prediction, anatomical correspondence, and temporal or paired-scan consistency.

The pretext task should not encourage smoothing away small vessels. Randomly masking large blocks may let the model infer lung texture without learning distal branches. Sampling strategies can emphasize tubular candidates, bifurcations, or regions close to the resolution limit. Cross-view prediction can require consistency among axial, coronal, and sagittal contexts. Paired contrast and non-contrast scans can teach intensity-invariant anatomy when registration quality is controlled.

Pretraining evaluation should separate representation quality from fine-tuning scale. Linear probes or frozen-encoder adapters can test what is already encoded; full fine-tuning can conceal weak pretraining

through task-specific learning. Comparisons should use equal labeled data and report performance curves as the number of annotated volumes decreases. The strongest case for a foundation model is not merely matching a specialist with full supervision, but retaining useful morphology when labels are scarce or domains change.

Priority 4: Redesign Prompts for Vascular Trees

SAM uses points and boxes because many natural-image objects are compact. A pulmonary tree crosses large parts of the volume, and one bounding box can include vessels, airways, nodules, and parenchymal abnormalities. Prompts should therefore express topology and anatomy. Candidate prompt types include central seed points, sparse centerline clicks, branch endpoints, lobe identity, vessel type, and exclusion points on adjacent airways or veins.

Interactive research should measure user effort, not only final accuracy. Count clicks, correction time, branch edits, and cognitive burden. A system that requires repeated distal prompting may be less useful than a slightly less accurate automatic model. Prompts can also be generated from another algorithm, but then the evaluation must include prompt-generation failures. An oracle prompt derived from the reference is not representative of clinical use.

Graph prompts are a promising direction. A user might confirm a central artery and one uncertain bifurcation, while the model propagates identity along connected branches. The network could predict alternative paths where evidence is ambiguous and ask for the most informative correction. This converts prompting from spatial hints into targeted topology repair. Uncertainty and expected error reduction should determine which branch is shown to the user.

Priority 5: Improve Labels Without Freezing Their Errors

Peripheral vessel annotation is expensive, and disagreement increases near the image-resolution limit. Foundation models can accelerate initial masks, but model-assisted annotation can create correlated errors if reviewers accept plausible false branches. The dataset should record provenance: manual from scratch, model-corrected, consensus, or adjudicated. A subset labeled independently by multiple experts is needed to estimate uncertainty and systematic omission.

Geometry-derived supervision inherits the mask. If a branch is missing, both its distance and thickness maps disappear. If a false bridge exists, skeleton propagation can turn it into a strong topological target. Research should therefore develop confidence-weighted priors. Stable central vessels can receive strong thickness supervision, while uncertain terminals receive softer or multi-rater targets. Morphological augmentation can simulate gaps, spurs, and caliber bias so the network learns to detect annotation defects rather than imitate them.

Weak labels may still be useful. Sparse centerlines, vesselness responses, or airway-vessel relationships can supplement incomplete masks. 3D U-Net showed that sparse annotations can train dense volumetric models when the learning setup is designed accordingly (Cicek et al., 2016). For foundation models, mixed supervision could combine a small set of complete masks with many partial centerlines and unlabeled CT scans. The evaluation must distinguish performance on visible-but-unlabeled branches from true false positives.

Priority 6: Integrate Topology With Foundation Representations

Topology should influence feature learning, not only final post-processing. clDice supplies a differentiable centerline objective for tubular structures (Shit et al., 2021). MorVess combines clDice with distance and thickness regression, but further work can model branch relationships explicitly. Graph neural modules can connect candidate centerline nodes; persistent-homology losses can penalize breaks and false cycles; differentiable path losses can protect routes from central trunks to terminals.

These constraints must remain image-conditioned. A loss that rewards connectivity too strongly may bridge nearby arteries and veins across low-contrast gaps. The optimal result is not the most connected possible mask but the most plausible tree supported by the scan. Local uncertainty, vessel orientation, and anatomical context should determine whether a gap is closed. Competing hypotheses can be retained for review instead of forcing a confident binary decision.

Caliber deserves equal attention. Abrupt diameter changes may be artifacts, but stenoses and pathological remodeling are real. A smoothness prior should not erase disease. The model can learn normal caliber transitions while allowing departures when image evidence is strong. Training and test sets must include vascular pathology so that an anatomical prior does not become a normality filter.

Priority 7: Evaluate Generalization as a Causal Question

External validation should identify why performance changes. Scanner vendor, dose, reconstruction, contrast, spacing, disease, and annotation convention are distinct causes. A single pooled external score cannot tell whether an adapter is robust or whether the new site happens to resemble the training distribution. Factorial or matched analyses can isolate acquisition and population effects.

The HiPaS work by Chu et al. (2025) demonstrates the scale of evidence possible in pulmonary artery-vein segmentation: pretraining on extensive chest CT, development on multi-center annotated data, external testing, comparison of non-contrast CT with CTPA, and downstream population analysis. It also shows that artery-vein identity and vessel abundance require richer evaluation than a binary vessel mask. Foundation-model research should include these downstream distinctions from the beginning.

PaSAL provides a complementary example. Eppink et al. (2026) combine an nnU-Net vessel segmentation with graph-based labeling of 19 vascular classes and test on an external clinical set. Their expert assessment reports that standard metrics correlate only weakly with perceived quality. This finding supports a translational principle: external evaluation should include anatomical coherence and correction burden, not only overlap.

MorVess performs cross-domain tests on pulmonary veins and airways, which supports the transferability of tubular priors. The next step is a multi-site study that holds the anatomical target constant while varying acquisition and disease. A prospective silent-deployment period could measure failure rates without affecting care. Only after that evidence should interactive or automated use be evaluated in a workflow.

Priority 8: Build Uncertainty and Review Into the Output

Foundation models can produce confident-looking masks even when the input is outside their training distribution. Calibration should be measured per voxel, but vascular review needs structural uncertainty. The system should estimate whether each branch exists, connects to the expected parent, and has a reliable caliber. Ensembles, stochastic adapters, test-time perturbations, or disagreement among mask, distance, and thickness heads can provide signals.

Uncertainty maps must be converted into actions. High uncertainty around a tiny terminal may be acceptable for gross vessel volume but unacceptable for surgical planning. The interface should use the declared clinical purpose to set review thresholds. It can prioritize a small number of questionable bifurcations and display alternative reconstructions. Every automatic correction or user edit should be logged so that recurrent failure modes inform later training.

Selective prediction is preferable to universal automation. The model can abstain on scans with extreme spacing, metal artifacts, severe motion, or unfamiliar postoperative anatomy. Evaluation should report coverage-risk curves: how performance changes as uncertain cases are referred. An apparently lower overall automation rate can produce a safer and more efficient system if expert time is concentrated where it matters.

Priority 9: Connect Segmentation to Clinical Measurements

Translation requires predefined downstream endpoints. For pulmonary hypertension research, relevant outputs might include central caliber, vascular volume, pruning, and right-to-left asymmetry. For resection planning, the identity and course of segmental vessels may dominate. For fibrosis, changes in small-vessel abundance and vessel-parenchyma relationships may be important. The model and evaluation should match the intended claim.

Derived measurements need uncertainty propagation. A segmentation ensemble can generate intervals for volume and caliber. Repeat scans and alternative reconstructions can test repeatability. Agreement with expert measurements should be assessed across the full range rather than summarized by correlation alone. A model can correlate well while having clinically important systematic bias.

Clinical studies should compare human-only, model-only, and human-with-model conditions where appropriate. Outcomes include accuracy, time, number of corrections, confidence, and error detection. Readers should be blinded to method identity when judging isolated outputs. Automation bias and overreliance should be measured, especially because three-dimensional renderings make structurally coherent errors persuasive.

A Staged Research Program

Stage 1 should reproduce the target method and strong specialist baselines on fixed public splits, reporting regional, surface, centerline, branch, caliber, memory, and latency outcomes. Reproduction should include preprocessing, physical spacing, geometric-target generation, and all ablations. Stage 2 should compare context mechanisms and parameter-efficient updates under matched budgets. Stage 3 should add multi-site, pathology-rich external data and analyze causal sources of domain shift.

Stage 4 should introduce interactive topology repair and branch-level uncertainty, measuring user effort with representative readers. Stage 5 should validate one predefined clinical or scientific measurement, including test-retest reliability and error propagation. Stage 6 should conduct silent prospective deployment, followed by a controlled workflow study if safety and performance thresholds are met.

At every stage, data splits, primary endpoints, and adaptation rules should be fixed before evaluation. Code, model configuration, and case-level metrics should be released when governance permits. Negative findings should be retained, because identifying where a 2.5D adapter fails or a geometric prior over-smooths pathology is more useful than another unqualified claim of generality.

Conclusion

Foundation models can contribute to pulmonary vessel segmentation, but only when their broad representations are paired with volumetric context, tubular geometry, and evidence aligned with vascular use. MorVess demonstrates a compelling combination: a frozen SAM encoder, lightweight 2.5D adaptation, multi-level fusion, joint mask-distance-thickness prediction, and staged training. Its reported efficiency and topology-aware results make it a valuable blueprint, while its short depth context and anisotropy limitations show why the problem is not solved. The next generation should allocate adaptation capacity anatomically, pretrain on volumetric structure, redesign prompts for trees, model uncertain topology, and validate downstream measurements across sites and diseases. Success will not be a foundation model that merely segments more voxels. It will be a system that reconstructs dependable vascular structure, knows when it is uncertain, and reduces the effort required to obtain clinically meaningful pulmonary measurements.

References

Mao, F., Chen, Y., Wu, B., Lin, L., Dai, J., Li, Z., ... & Qin, F. (2026). MorVess: Morphology-Aware Pulmonary Vessel Segmentation Network. *arXiv preprint arXiv:2606.24214*.

Chu, Y., Luo, G., Zhou, L., Cao, S., Ma, G., Meng, X., ... & Gao, X. (2025). Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences. *Nature Communications*, *16*, 2262.

Cicek, O., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016* (pp. 424-432). Springer.

Eppink, J., Kervadec, H., van Capelleveen, J., Verhoeff, J., Senan, S., & Bohoudi, O. (2026). PaSAL: A deep learning pipeline for pulmonary artery-vein segmentation and anatomical labeling in thoracic CT. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning* (pp. 3561-3592). PMLR.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2022). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (pp. 272-284). Springer.

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16000-16009).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*(2), 203-211.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015-4026).

Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., ... & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, *15*, 654.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015* (pp. 234-241). Springer.

Shit, S., Paetzold, J. C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., ... & Menze, B. H. (2021). clDice - A novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16560-16569).

Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., ... & He, J. (2025). SAM-Med3D: A vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(10), 17599-17612.