# Highlights

## FedSAM3D: Auto-Prompted Volumetric Multi-Organ Segmentation with Coupled Distillation for Federated Learning under Label Fragmentation

Hamza Mukhtar

- FedSAM3D enables federated 3D multi-organ segmentation with non-IID + partial labels.
- Our auto prompt generator makes SAM-style 3D segmentation fully automatic.
- Masked supervision avoids “missing class = background” and reduces class-collision errors.
- Supervision-mass aggregation stabilizes fusion when label coverage varies across sites.
- Coupled global+peer distillation transfers missing-organ knowledge with confidence gating.

# FedSAM3D: Auto-Prompted Volumetric Multi-Organ Segmentation with Coupled Distillation for Federated Learning under Label Fragmentation

Hamza Mukhtar[a]

[a]Computer Science department, University of Engineering and Technology, Lahore, 53605, Pakistan

ARTICLE INFO

*Keywords*:
Medical Imaging
Deep Learning
Transformer
Federated Learning
Knowledge Distillation
Volumetric Adaptation
Segment Anything Model
Prompt-conditioned Segmentation

ABSTRACT

Multi-organ segmentation in volumetric CT/MRI is essential for organ volumetry, radiotherapy contouring, and surgical planning, yet clinical deployment rarely matches the centralised-learning assumption of pooled data and dense multi-organ annotation. In practice, imaging cohorts remain privacy-isolated across institutions and exhibit pronounced non-IID shifts (scanner, protocol, population), while annotations are typically fragmented (single-organ or study-specific label sets). This creates a federated 3D segmentation regime with simultaneous optimization instability and label-space mismatch, where naïve objectives treat missing organs as background and induce systematic false-negative gradients, class collision, and degraded global generalization. Existing federated partial-label methods largely target conventional encoder–decoder models, whereas SAM-style promptable models are predominantly 2D-centric and are not designed for volumetric coherence, heterogeneous supervision, or stable federated optimization. To address this gap, we propose a federated framework (FedSAM3D) that learns a single prompt-conditioned volumetric segmenter from heterogeneous, partially labeled clients. FedSAM3D (i) introduces parameter-efficient 3D adaptation of a SAM-style backbone (depth-aware tokenization/positional encoding and lightweight volumetric adapters) to reduce communication and mitigate client drift, (ii) replaces manual interaction with an Auto Prompt Generator that synthesizes class-conditioned query embeddings for fully automatic multi-organ inference, (iii) enforces partial-label correctness via masked supervision that updates only locally annotated organs, and (iv) stabilizes cross-silo training with supervision-mass aggregation plus a coupled knowledge-transfer scheme comprising confidence-gated global-consistency distillation and peer-guided distillation using lightweight organ experts with stochastic missing-class subsampling. Evaluated on a cross-dataset federation of five abdominal CT benchmarks with fragmented organ annotations, FedSAM3D achieves high performance across all clients: Dice scores of 0.961 (spleen), 0.973 (liver), 0.862 (pancreas), 0.956 (kidney), and 0.851 macro-average across 13 BTCV organs, with corresponding HD95 distances of 2.9, 3.5, 8.8, 3.8, and 6.6 mm. The framework maintains a communication payload of only 12.4 MB/round-7× smaller than a distillation-based mechanism which makes robust multi-organ segmentation under realistic privacy and supervision constraints.

## 1. Introduction

Accurate multi-organ segmentation in volumetric CT and MRI is a foundational capability for organ volumetry, surgical and interventional planning, radiotherapy contouring, and downstream lesion analysis [9]. In centralized settings, self-configuring pipelines and volumetric transformer architectures (e.g., nnU-Net and UNETR) have achieved strong performance by leveraging pooled training cohorts and dense multi-organ supervision [1]. In routine practice, however, these assumptions are frequently violated: (i) clinical imaging data are distributed across institutions and cannot be centralized due to privacy, governance, and compliance constraints, and (ii) voxel-wise annotations are expensive and typically protocol-driven, such that many sites label only a subset of clinically relevant structures.

Federated learning (FL) enables privacy-preserving cross-institution collaboration by optimizing a shared model without exchanging raw patient data [32; 35; 10]. Recent segmentation-oriented FL pipelines further indicate that careful system design can yield robust multi-hospital models [38]. Nonetheless, federated 3D segmentation remains difficult due to statistical heterogeneity and optimization instability: cross-site differences in scanner vendors, acquisition protocols, reconstruction parameters, and patient cohorts induce non-IID feature distributions, while local optimization can cause client drift that harms global generalization [24; 17; 25]. These challenges are exacerbated in volumetric settings where GPU memory constraints enforce patch-based training and sliding-window inference, increasing gradient variance and amplifying sensitivity to domain shift during aggregation.

A second and often dominant barrier is supervision heterogeneity (Figure 1). In multi-organ segmentation, institutions frequently annotate only the organs relevant to local studies (e.g., single-organ cohortresulting in partial-label and label-space mismatch regimes in which unlabeled organs are present but must not be treated as background. Naive objectives introduce systematic false-negative gradients for missing classes and can induce class collision during federated fusion. Recent work addresses partially labeled federated segmentation through unified label learning,

*Corresponding author
hamza.hm.mukhtar@gmail.com (H. Mukhtar)
ORCID(s):

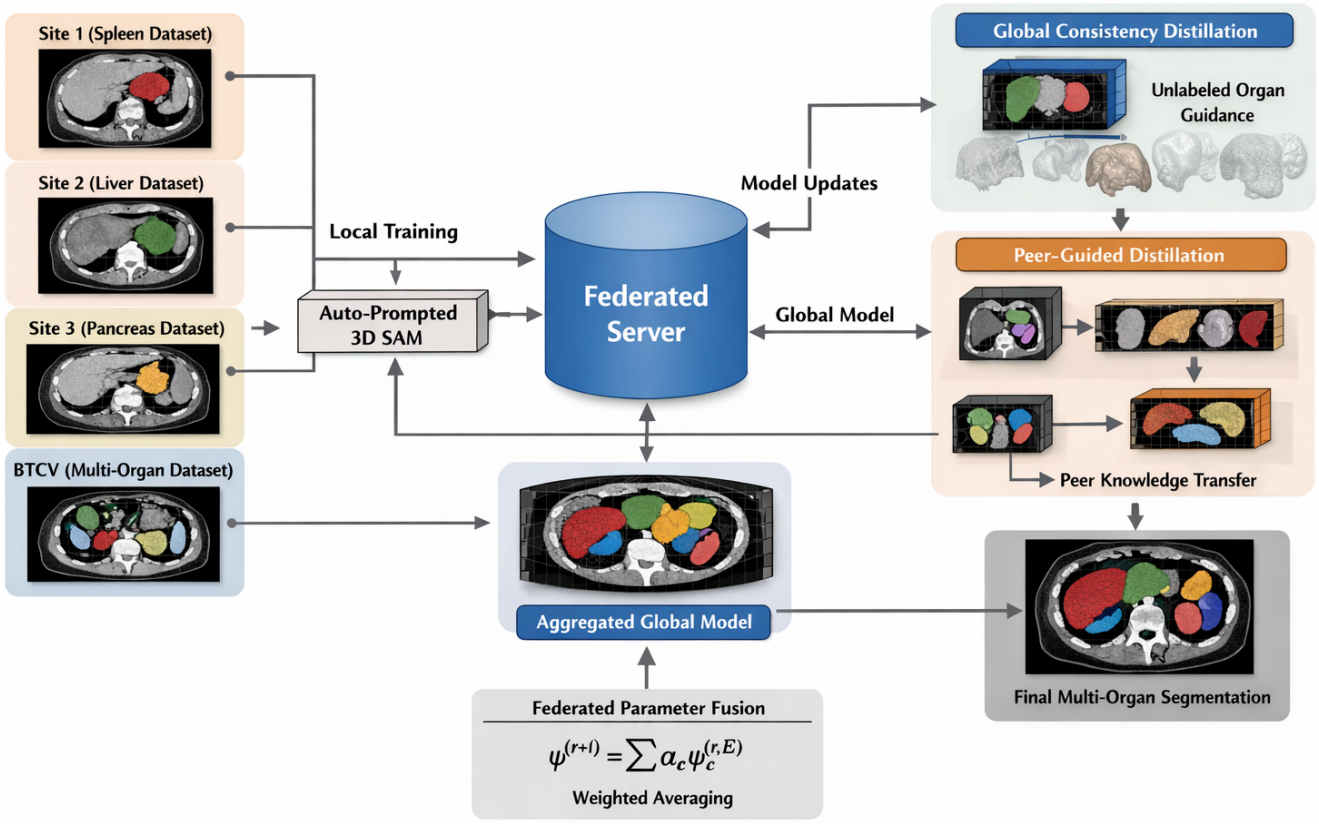


**Figure 1:** Schematic of the proposed FedSAM3D for multi-class 3D medical image segmentation under partial-label supervision. Multiple privacy-isolated sites hold heterogeneous datasets with disjoint/overlapping organ annotations (e.g., single-organ cohorts and a multi-organ cohort) and perform local optimization of an auto-prompted 3D SAM model using a masked objective that updates only classes annotated at each site, avoiding "missing class = background" supervision. Clients transmit model updates to a central federated server, which conaggregates aobal model via supervision-weighted parameter fusion. To mitigate label-space mismatch and non-IID drift, the global model provides global-consistency distillation as soft guidance for locally unlabeled organs, while peer-guided distillation transfers organ-specific expertise via lightweight peer experts.

completeness-aware reweighting, or expert-style decompositions [15; 43; 44]. Distillation-based FL further shows promise for transferring missing-class knowledge and reducing communication overhead [21; 19]. However, most existing solutions are developed for conventional encoder-decoder segmenters and do not directly account for prompt-conditioned, foundation-model-based segmentation in 3D under realistic cross-silo non-IID shifts and fragmented supervision.

Promptable segmentation foundation models offer a complementary paradigm by decoupling target specification from mask generation. Segment Anything (SAM) introduced prompt-conditioned mask decoding with strong transferability across objects and datasets [22], and medical-domain scaling and parameter-efficient tuning have substantially improved SAM-style performance under modality shift [30; 42]. Yet, most SAM-based systems remain primarily 2D-centric and require repeated prompting to segment volumetric scans, which is impractical for routine multi-class workflows. Recent efforts mitigate this interaction burden through automatic prompting and slice-to-volume propagation [26; 18; 46], but they typically assume centralized training and do not consider federated optimization where both imaging distributions and label availability differ across sites. Consequently, an open gap remains at the intersection of (i) volumetric prompt-conditioned decoding, (ii) partial-label multi-organ supervision across clients, (iii) parameter-efficient communication under large 3D backbones, and (iv) stability under non-IID federated optimization.

These limitations reveal a significant research gap at the intersection of three dimensions: (i) intrinsically volumetric prompt-conditioned decoding that preserves 3D anatomical coherence, (ii) robust handling of supervision heterogeneity where label spaces are fragmented across sites, and (iii) federated optimization stability under combined non-IID imaging distributions and partial-label regimes. Existing works address subsets of these challenges in isolation—volumetric SAM adaptations ignore federation constraints, FL methods for partial labels neglect foundation model capabilities,

and distillation techniques lack organ-specific specialization transfer, but none provide an integrated solution for realistic cross-institutional multi-organ segmentation.

To bridge this gap, we propose a federated framework for learning a single prompt-conditioned 3D segmenter (FedSAM3D) from privacy-isolated, non-IID datasets with fragmented organ annotations (Figure 1). FedSAM3D adapts a SAM-style backbone to volumetric CT/MRI using parameter-efficient 3D components (e.g., factorized volumetric tokenization, separable 3D positional encoding, and lightweight 3D adapters) while keeping the majority of pretrained parameters frozen, thereby reducing communication payloads and mitigating drift. To eliminate manual interaction and enable scalable multi-organ training, an Auto Prompt Generator synthesizes class-conditioned query embeddings directly from volumetric features, supporting fully automatic multi-class inference. Critically, FedSAM3D is partial-label aware: supervised losses are masked to include only locally annotated organs, preventing the "missing class = background" failure mode. Moreover, because label fragmentation can underrepresent rare or site-specific organ expertise in the global model, we introduce a coupled knowledge-transfer strategy with two complementary pathways. First, global consistency distillation regularizes each client on locally unlabeled organs using confidence-gated soft targets produced by the current server model, stabilizing learning and reducing forgetting under non-IID shift. Second, peer-guided distillation transfers organ-specialised knowledge via lightweight class-expert modules extracted from clients and relayed across the federation; peer predictions are ensembled and confidence-filtered to suppress unreliable guidance while preserving scalability through stochastic subsampling of missing classes. Finally, to account for heterogeneous label coverage, FedSAM3D employs supervision-aware aggregation that weights client updates by their effective supervised signal, improving stability relative to size-only or uniform weighting when annotation density varies across sites.

Our main contributions are:

- We propose a federated learning framework (FedSAM3D) for prompt-conditioned 3D multi-organ segmentation under realistic cross-silo non-IID imaging and fragmented organ supervision.
- We introduce parameter-efficient volumetric adaptation of a SAM-style backbone, coupled with an Auto Prompt Generator that enables fully automatic class-conditioned prompting for scalable multi-organ training and inference.
- We develop a partial-label-aware federated objective with masked supervision and supervision-aware parameter fusion to prevent "missing class = background" gradients and reduce bias under heterogeneous label coverage.
- We propose a coupled global-peer distillation mechanism with confidence gating, lightweight peer experts, and stochastic missing-class subsampling to transfer missing-class knowledge efficiently and preserve organ competence across heterogeneous clients.

## 2. Literature Review

### 2.1. Federated 3D Segmentation under Non-IID Imaging

Federated learning (FL) enables privacy-preserving multi-institution training in medical imaging, but volumetric memory constraints and cross-site distribution shifts can degrade convergence and generalization [10]. While centralized 3D segmenters such as nnU-Net and volumetric transformers (e.g., UNETR) achieve strong performance, their federated counterparts must address optimization instability and site heterogeneity.

Most early federated segmentation methods rely on FedAvg [32], which is communication efficient yet prone to client drift under non-IID data. Heterogeneity-aware optimizers, including proximal regularization (FedProx) [24] and control-variate correction (SCAFFOLD) [17], improve stability, while normalization-based personalization (FedBN) mitigates feature-shift heterogeneity by keeping client-specific normalization statistics local [25]. Recent segmentation-tailored FL systems further refine aggregation via reliability-aware weighting and uncertainty/gradient-based fusion, improving robustness to cross-domain variability and small-target imbalance.

A second heterogeneity axis is label-space inconsistency and partial annotation, where institutions label only organ subsets; naively treating missing classes as background induces false negatives and destabilizes training. UFPS introduces unified label learning with sharpness-aware optimization for partially supervised federated segmentation [15], and FedIA models incomplete annotations as low-quality supervision with completeness-aware reweighting and correction [43]. Finally, foundation models are beginning to influence federated segmentation: SAM provides promptable transferability [22], and FedFMS explores federated adaptation of SAM-style models with medical adapters for improved communication and training efficiency [29]. Nevertheless, robust 3D prompt-conditioned segmentation under realistic non-IID shifts and partial-label multi-organ supervision remains underexplored, motivating frameworks such as FedSAM3D.

### 2.2. Multi-Organ Partial-Label Mismatch

Multi-organ medical image segmentation is commonly studied under dense supervision, where all target structures are labeled in every scan. In routine clinical practice, however, annotations are typically fragmented across institutions and studies, yielding a partial-label regime in which the union of datasets spans a large organ vocabulary, but each dataset supervises only a small subset. A key difficulty is that unlabeled organs are often present in the image yet lack ground truth; treating these voxels as background induces systematic false-negative gradients and biased decision boundaries.

To mitigate "missing = background" collapse under incomplete labels, prior work has explored task-conditioned formulations (e.g., conditioning on organ identity) and multi-head decoders, as well as semi-supervised/self-training with pseudo-labels. However, pseudo-labelling can be unstable when labelled and unlabeled regions exhibit mismatched feature distributions, especially for low-contrast organs. Recent Medical Image Analysis work explicitly analyses this mismatch and proposes distribution-alignment strategies (e.g., cross-set mixing and prototype-based alignment) for partially supervised multi-organ learning [16].

Partial-label challenges are amplified in federated learning, where clients may annotate distinct, potentially non-overlapping organ subsets, creating abel-space mismatch. In this setting, naive aggregation can cause class collision and inconsistent gradients for absent classes. Fed-MENU addresses this via a multi-encoding expert design that enables federated multi-organ training without label alignment [44]. UFPS formalizes federated partially supervised segmentation and introduces unified label learning with sharpness-aware optimization to reduce class collision and client drift [15], while FedIA models incomplete annotations as low-quality supervision and incorporates completeness-aware reweighting and correction to stabilize training [43]. In practical deployments with additional unlabeled data, FPS-Seg unifies partial supervision, semi-supervision, and FL to exploit distributed partially labeled and unlabeled CT scans within a single pipeline [45]. More recently, distillation-based FL has been proposed to reconcile heterogeneous label spaces and reduce drift; in particular, a few-round strategy distils locally pretrained models into a global model using public data augmented with server-side synthesis to mitigate proxy data shift [19]. Despite these advances, most existing solutions are developed for conventional encoder-decoder architectures and do not directly target prompt-conditioned, foundation-model-based 3D segmenters.

## 2.3. Federated Distillation for Segmentation

Knowledge distillation (KD) has progressed from model compression to a general mechanism for transferring structured information in dense prediction, where supervision must preserve geometry, boundaries, and inter-class relations. Beyond response-level logit matching, segmentation benefits from structure-aware distillation, including feature alignment, relational/affinity transfer, and boundary-sensitive objectives; for instance, BPKD separates edge and body knowledge to improve calibration and contour fidelity [28].

Promptable/foundation segmenters further broaden KD's utility in medical imaging, where domain shift and limited supervision often necessitate teacher-student transfer. DES-SAM exploits SAM-driven pseudo-labelling and self-distillation prompting for box-supervised semantic segmentation [13], while UM-SAM combines pseudo-label, feature, and contrastive distillation to transfer SAM knowledge to lightweight models under unsupervised adaptation [9].

In federated learning (FL), distillation complements or replaces parameter averaging when data are non-IID, communication is heavy, or client architectures differ. Server-side ensemble distillation refines a global model by imitating an ensemble of client predictors on proxy data; FedDF is a representative approach that improves fusion and relaxes strict homogeneity assumptions [27]. An IJCAI survey organizes KD-based FL into server-side refinement, client-side local-global regularization, and generator/data-free distillation, highlighting trade-offs in communication, privacy exposure, and robustness [34].

Distillation-based FL has been specialized for medical segmentation, where partial labels and organ imbalance exacerbate drift under heterogeneity. Kim *et al.* propose KD-regularized FL for multi-organ segmentation from partially labeled datasets by combining global guidance with organ-specific teachers to reduce forgetting [21]. Subsequently, a communication-efficient framework distils client knowledge into a global model and uses image synthesis to mitigate proxy-data mismatch, enabling effective few-round training [20]. Despite these advances, KD designs that jointly handle 3D dense prediction, prompt-conditioned decoding, and partial-label multi-organ supervision under realistic cross-silo non-IID shifts remain limited, motivating coupled global-local distillation strategies as in FedSAM3D.

## 2.4. 3D Adaptation of SAM-Style Segmenters

Segmentation foundation models have reframed interactive and open-vocabulary delineation by decoupling target specification (prompts) from mask generation. Segment Anything (SAM) combines a high-capacity image encoder, a prompt encoder (e.g., points/boxes/masks), and a mask decoder that fuses visual and prompt embeddings [22]. Direct transfer to medical imaging is challenged by modality shift and anatomy-specific constraints, motivating medical-domain scaling and parameter-efficient adaptation. MedSAM improves promptable generalization via large-scale medical supervision [30], while adapter-style tuning updates a small parameter subset to reduce compute/memory while retaining pretrained representations [42]. Prompt adaptation has also been explored for robustness under protocol shift, e.g., domain-adaptive prompting in DAPSAM [41].

A central limitation of 2D promptable segmenters is volumetric mismatch: slice-wise use requires repeated prompting and can lead to interslice inconsistencies. Existing 3D extensions follow two main directions. 3D-by-aggregation pipelines retain a 2D backbone and propagate sparse cues across slices (e.g., RadSAM) [18]. Conversely, intrinsically volumetric promptable models adapt SAM-style components to 3D inputs, often coupled with automatic prompting to reduce interaction (e.g., 3D-SAutoMed) [26]. However, volumetric prompting increases compute and memory demands, motivating efficiency-oriented prompting pipelines and universal volumetric segmenters (e.g., AutoProSAM and SegVol). Collectively, these efforts highlight an open gap at the intersection of volumetric prompt-conditioned decoding, automatic prompt synthesis, and robustness under

heterogeneous data and incomplete supervision, motivating federated designs that adapt SAM-style 3D segmenters while controlling communication and stabilizing learning under partial labels.

## 3. Proposed Framework

### 3.1. Problem Formulation

We consider a cross-silo federated setting with $K$ privacy-isolated institutions (clients). Client $k$ holds a local dataset $\mathcal{S}_k$ of $N_k$ volumetric scans (CT/MRI) and corresponding voxel-wise annotations for only a subset of the global organ vocabulary. Due to differences in scanners, acquisition protocols, reconstruction pipelines, and patient cohorts, the client data distributions are generally non-identically distributed (non-IID).

Let $\mathcal{O} = \{1, \dots, C\}$ denote the global set of target organs and $\mathcal{O}_k \subseteq \mathcal{O}$ the subset annotated at client $k$. For organs $c \in \mathcal{O}_k$, the client provides a binary mask $\mathbf{G}_{k,n}^{(c)}$ for scan $n$. A central requirement in partial-label learning is that organs *not* annotated at a client must not be implicitly treated as background. We therefore use a per-client class-availability indicator $\mathbb{I}_k(c)$, where $\mathbb{I}_k(c) = 1$ if $c \in \mathcal{O}_k$ and 0 otherwise.

FedSAM3D instantiates a single prompt-conditioned 3D segmenter that outputs a per-organ probability volume for each class $c$, conditioned on automatically generated prompts (Sec. 3.8). In training, supervision is applied *only* to organs annotated at the current client. Concretely, the masked supervised loss at client $k$ is computed by summing a standard volumetric segmentation criterion (Dice + voxel-wise classification loss) over the locally labeled organs:

$$\mathcal{L}_k^{\text{sup}} = \sum_{n=1}^{N_k} \sum_{c=1}^{C} \mathbb{I}_k(c)\, \mathcal{E}\Big(\widehat{\mathbf{P}}_{k,n}^{(c)}, \mathbf{G}_{k,n}^{(c)}\Big), \tag{1}$$

where $\widehat{\mathbf{P}}_{k,n}^{(c)}$ is the predicted probability map for organ $c$ and $\mathcal{E}(\cdot)$ denotes a Dice–cross-entropy (or Dice–BCE) composite computed voxel-wise. This masking eliminates the "missing class = background" failure mode while preserving dense supervision for labeled organs.

### 3.2. Global Consistency Distillation

In partial-label federated multi-organ segmentation, each site is supervised on only a subset of organs, which can cause the local learner to specialize toward its annotated targets and gradually lose competence on unobserved categories. This phenomenon is amplified under non-IID imaging distributions and leads to unstable global fusion when only supervised losses are used. To counteract this drift, FedSAM3D introduces a global consistency distillation mechanism, as shown in Figure 2, in which the server model provides soft guidance on organs that are not annotated at a given site. In effect, each client benefits from the federation-wide organ knowledge even when its local dataset lacks corresponding labels, improving cross-site coherence and reducing forgetting.

At round $r$, the current server model $f_{\boldsymbol{\psi}^{(r)}}$ acts as a teacher and produces probabilistic organ maps for a client volume $\mathbf{V}_{k,n}$. The client model $f_{\boldsymbol{\psi}_k}$ (initialized from $\boldsymbol{\psi}^{(r)}$) serves as the student. Distillation is applied only to the organ set that is missing supervision at site $k$, i.e., $\overline{\mathcal{O}}_k = \mathcal{O} \setminus \mathcal{O}_k$, so that no distillation term interferes with true labels. Let $\widehat{\mathbf{P}}_{k,n}^{(c)}$ and $\widehat{\mathbf{Q}}_{k,n}^{(c)}$ denote the student and teacher probability maps for class $c$, respectively.

Direct distillation across all voxels and all missing classes can propagate low-confidence teacher errors. To mitigate this shortcoming, we employ a confidence-gated mask that selects reliable teacher supervision. Specifically, define:

$$\mathbf{M}_{k,n}(v) = \mathbb{1}\left[\max_{c \in \overline{\mathcal{O}}_k} \widehat{\mathbf{Q}}_{k,n}^{(c)}(v) \geq \tau\right], \tag{2}$$

where $\tau \in (0, 1)$ is a confidence threshold and $\mathbb{1}[\cdot]$ is the indicator function. Using the voxel lattice $\Omega_{k,n}$ with $|\Omega_{k,n}| = DHW$, the global consistency distillation loss is written as a masked cross-entropy between teacher and student soft targets:

$$\mathcal{E}_{k,n}^{\text{GKD}} = \frac{1}{|\Omega_{k,n}||\overline{\mathcal{O}}_k|} \sum_{c \in \overline{\mathcal{O}}_k} \sum_{v \in \Omega_{k,n}} \mathbf{M}_{k,n}(v)\ \text{BCE}_\epsilon\big(\widehat{\mathbf{P}}_{k,n}^{(c)}(v), \widehat{\mathbf{Q}}_{k,n}^{(c)}(v)\big). \tag{3}$$

Compared with an ungated formulation, it reduces the impact of uncertain teacher predictions and yields more stable optimization when label coverage is sparse or when imaging shifts are large. The overall client-side objective augments the masked supervised segmentation energy with the global distillation term:

$$\mathcal{J}_k = \sum_{n=1}^{N_k} \left[\sum_{c=1}^{C} \mathbb{I}_k(c)\, \mathcal{E}\Big(\widehat{\mathbf{P}}_{k,n}^{(c)}, \mathbf{G}_{k,n}^{(c)}\Big) \;+\; \beta\, \mathcal{E}_{k,n}^{\text{GKD}}\right], \tag{4}$$

where $\beta > 0$ controls the strength of global guidance. This composite objective ensures that (i) annotated organs are learned from ground truth without contamination from missing labels, and (ii) unannotated organs are regularized toward federation-wide predictions, improving global alignment and preserving multi-organ competence under privacy constraints [33].

### 3.3. Peer-Guided Distillation

While global consistency distillation transfers federation-wide knowledge from the server model to each site, it can under-represent rare or site-specific organ expertise because the server teacher reflects an averaged consensus. To complement this, FedSAM3D incorporates *peer-guided distillation* (PKD), where clients share lightweight organ-specific expert modules that provide soft guidance for organs missing local supervision (Figure 3). This transfer is performed without exchanging raw images and is especially effective under fragmented organ annotations across sites [33].

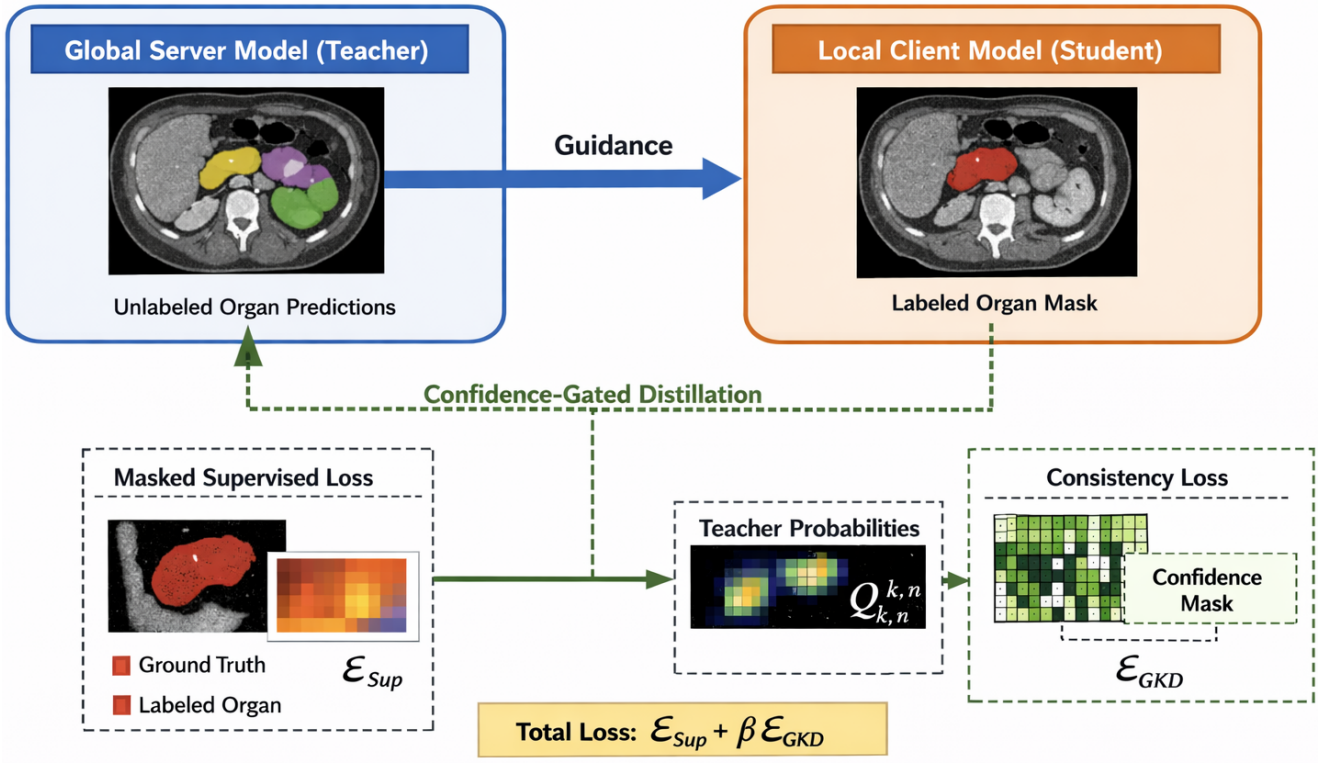


**Figure 2:** Global consistency distillation in FedSAM3D. The federated server model (teacher) generates soft predictions for organs unlabeled at a client, which are transferred to the local client model (student) via confidence-gated distillation to suppress low-confidence guidance. Training minimizes a composite objective comprising masked supervised loss on locally annotated organs, $\mathcal{E}_{\text{sup}}$, and a gated consistency loss, $\mathcal{E}_{\text{GKD}}$, weighted by $\beta$: $\mathcal{E}_{\text{sup}} + \beta\,\mathcal{E}_{\text{GKD}}$.

A direct local-KD approach that trains and exchanges a full organ-specific model per class is computationally prohibitive. Instead, each client $k$ maintains a small set of parameter-efficient class experts $\{\boldsymbol{\omega}_k^{(c)}\}_{c\in\mathcal{O}_k}$ (e.g., lightweight adapters/heads) extracted from its locally trained FedSAM3D model. After local training at round $r$, the client uploads only these experts to the server, which relays them to other participants. For a target client $k$ and a missing organ $c \in \overline{\mathcal{O}}_k = \mathcal{O} \setminus \mathcal{O}_k$, define the set of providers $\mathcal{P}_k(c) = \{p \neq k \mid c \in \mathcal{O}_p\}$. Each provider expert produces a peer probability map on the target client's volume $\mathbf{V}_{k,n}$ (with auto-prompts $\boldsymbol{\pi}_{k,n}$) as $\widehat{\mathbf{Q}}^{(c)}_{p\to k,n} = g_{\boldsymbol{\omega}_p^{(c)}}(\mathbf{V}_{k,n}, \boldsymbol{\pi}_{k,n})$. When multiple providers exist, we ensemble their outputs and apply confidence gating to suppress unreliable guidance:

$$\begin{aligned} \bar{\mathbf{Q}}^{(c)}_{k,n} &= \sum_{p\in\mathcal{P}_k(c)} \eta_p^{(c)}\, \widehat{\mathbf{Q}}^{(c)}_{p\to k,n}, \\ \mathbf{M}^{\mathrm{P}}_{k,n}(v) &= \mathbb{1}\left[\max_{c\in\widetilde{\mathcal{O}}_k^{(r)}} \bar{\mathbf{Q}}^{(c)}_{k,n}(v) \geq \tau_{\mathrm{P}}\right]. \end{aligned} \tag{5}$$

where $\eta_p^{(c)}$ are convex weights (uniform or reliability-weighted), $\tau_{\mathrm{P}} \in (0,1)$ is a peer-confidence threshold, and $\widetilde{\mathcal{O}}_k^{(r)} \subseteq \overline{\mathcal{O}}_k$ denotes the subset of missing classes selected for PKD in round $r$.

Given the student prediction $\widehat{\mathbf{P}}^{(c)}_{k,n}$ for organ $c$, PKD minimizes a confidence-masked cross-entropy between the peer ensemble target and the student output:

$$\mathcal{E}^{\mathrm{PKD}}_{k,n} = \mathbb{E}_{c\sim\mathcal{U}(\widetilde{\mathcal{O}}_k^{(r)}),\, v\sim\mathcal{U}(\Omega_{k,n})}\left[\mathbf{M}^{\mathrm{P}}_{k,n}(v) - \sum_i \bar{\mathbf{Q}}^{(c)}_{k,n}(v)_i \log(\widehat{\mathbf{P}}^{(c)}_{k,n}(v)_i + \epsilon)\right]. \tag{6}$$

where $\Omega_{k,n}$ is the voxel lattice and $-\sum_i \cdot_i \log(\cdot_i + \epsilon)$ denotes cross-entropy between soft targets and probabilities (defined in the preamble). This formulation transfers organ-specific competence from clients with ground-truth supervision for organ $c$, while filtering out low-confidence voxels prone to domain-shift errors.

To keep peer transfer scalable when $|\overline{\mathcal{O}}_k|$ is large (e.g., single-organ clients), we perform stochastic missing-class subsampling: at each round (or minibatch), client $k$ samples a small set $\widetilde{\mathcal{O}}_k^{(r)} \subseteq \overline{\mathcal{O}}_k$ with $|\widetilde{\mathcal{O}}_k^{(r)}| = m_k \ll |\overline{\mathcal{O}}_k|$. This yields a Monte Carlo estimate of the full PKD objective, substantially reducing computational and memory costs while preserving the benefits of peer specialization.

Finally, combining masked supervised learning, global consistency distillation, and peer-guided distillation yields the per-sample objective:

$$\mathcal{L}_{k,n} = \sum_{c=1}^{C} \mathbb{I}_k(c)\, \mathcal{E}\left(\widehat{\mathbf{P}}^{(c)}_{k,n}, \mathbf{G}^{(c)}_{k,n}\right) + \beta\, \mathcal{E}^{\mathrm{GKD}}_{k,n} + \gamma\, \mathcal{E}^{\mathrm{PKD}}_{k,n}, \tag{7}$$

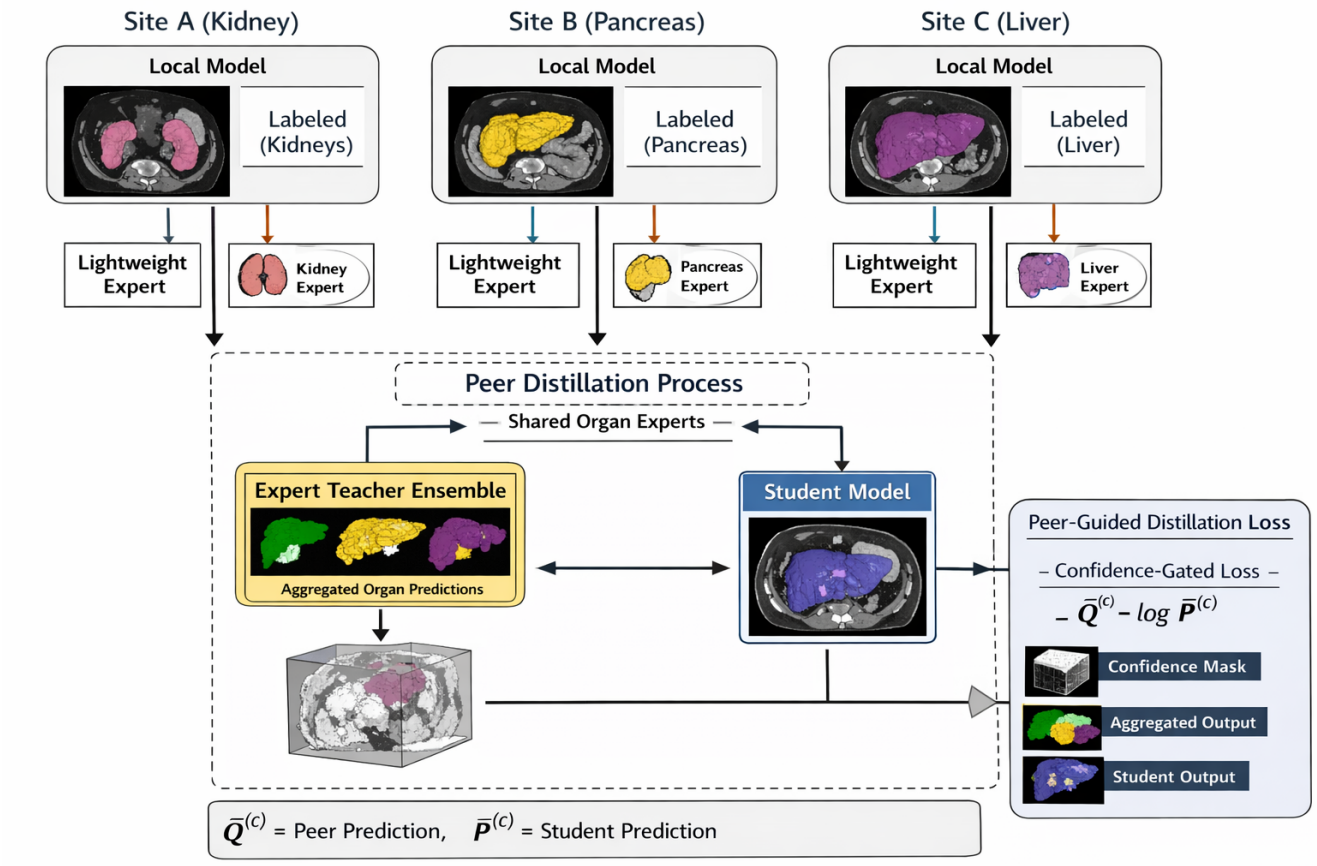


**Figure 3:** Peer-guided distillation in FedSAM3D for partial-label federated multi-organ segmentation. Each site trains a local model on its annotated organ subset under supervision and extracts lightweight, organ-specific expert modules (e.g., kidney/pancreas/liver experts). These experts are shared across sites and form a peer teacher ensemble that produces aggregated soft predictions for organs missing local annotations at a target client. The target client (student) is then regularized to match the peer ensemble outputs via a confidence-gated distillation objective, using a voxel-wise confidence mask to suppress unreliable peer guidance.

and the client objective $\mathcal{J}_k = \sum_{n=1}^{N_k} \mathcal{L}_{k,n}$. This design preserves privacy (no raw data exchange), improves missing-organ competence under partial labels, and remains computationally feasible by restricting peer transfer to lightweight class experts and a sampled subset of missing classes.

### 3.4. Coupled Global-Peer Distillation Strategy

Federated multi-organ segmentation under partial labels is hindered by three interacting factors: (i) domain shift across sites (non-IID imaging), (ii) incomplete supervision per site (fragmented organ annotations), and (iii) client drift, where local optimization overemphasizes annotated targets and degrades performance on missing classes. FedSAM3D addresses these limitations by jointly deploying (a) server-to-client consistency guidance and (b) peer specialization transfer. Their combination provides complementary supervision signals: the former enforces federation-wide semantic alignment, while the latter injects fine-grained organ expertise that may be diluted by global averaging.

The global distillation encourages each client model to remain compatible with the server model on the set of locally unannotated organs $\overline{\mathcal{O}}_k$. This alignment acts as a regularizer against forgetting by constraining the local solution space to preserve organ knowledge accrued across the federation, even when those organs are absent from local annotations. In addition, the confidence gating in (2) mitigates a key drawback of naive distillation: blindly propagating uncertain teacher predictions on out-of-domain voxels.

In parallel, peer-guided distillation enables each client to exploit organ-specific expertise from other sites through lightweight class experts (or compressed soft guidance maps). This decentralized knowledge flow is especially valuable when certain organs are well annotated only at a subset of sites or when the global model underfits rare structures due to averaging effects. To keep the approach scalable, FedSAM3D restricts peer transfer to (i) parameter-efficient experts rather than full organ-specific networks and (ii) stochastically sampled missing classes per round, avoiding quadratic growth in compute and memory with the organ vocabulary size.

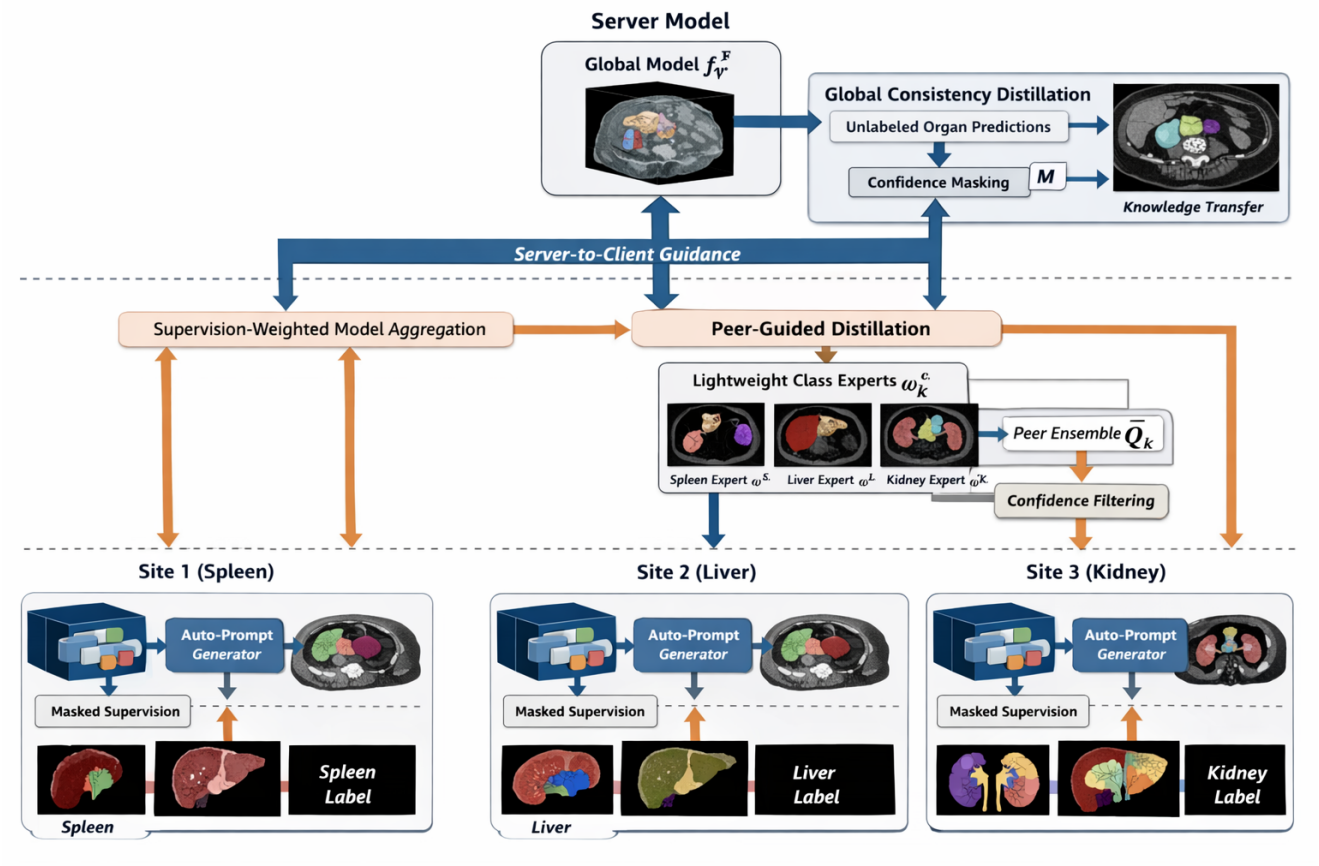


**Figure 4:** Coupled global-peer distillation in FedSAM3D. Each client trains an auto-prompted 3D SAM model with masked supervision, i.e., supervised losses are computed only for locally annotated organs. Client updates are fused via supervision-weighted aggregation to form the global server model $f^F$. Two complementary transfer pathways mitigate label-space mismatch and non-IID drift: (i) global consistency distillation, where $f^F$ provides confidence-masked soft guidance for locally unlabeled organs, and (ii) peer-guided distillation, where lightweight class experts $\omega_k^c$ are shared and ensembled to produce peer predictions $\bar{Q}_k$ with confidence filtering.

As shown in Figure 4, coupling both mechanisms yields a unified local objective that combines (1) masked supervised segmentation on locally labeled organs, (2) server-guided distillation on missing organs, and (3) peer-guided distillation on a sampled subset of missing organs:

$$\mathcal{J}_k = \sum_{n=1}^{N_k} \mathcal{L}_{k,n}, \tag{8}$$

$$\mathcal{L}_{k,n} = \underbrace{\sum_{c=1}^{C} \mathbb{I}_k(c)\, \mathcal{E}\left(\widehat{\mathbf{P}}_{k,n}^{(c)}, \mathbf{G}_{k,n}^{(c)}\right)}_{\text{masked sup.}} + \beta \underbrace{\mathcal{E}_{k,n}^{\text{GKD}}}_{\text{server cons.}} + \gamma \underbrace{\mathcal{E}_{k,n}^{\text{PKD}}}_{\text{peer transf.}}. \tag{9}$$

where $\beta$ and $\gamma$ control the relative influence of the two distillation pathways. In practice, we recommend scheduling these weights (e.g., warm up $\beta$ early for stable semantic alignment, then gradually increase $\gamma$ as peer experts become reliable), which addresses another common shortcoming: overly strong distillation in early rounds can lock the optimization into a suboptimal consensus before organ-specific competence emerges. The coupled strategy can be interpreted as coarse-to-fine regularization. Global consistency distillation provides a coarse, federation-level prior that stabilizes training under non-IID data and partial labels, while peer-guided distillation supplies fine, organ-targeted corrections that improve delineation of underrepresented or anatomically complex structures. Together, they promote uniform propagation of shared knowledge and targeted refinement where the global model alone is insufficient, improving segmentation robustness across heterogeneous sites without exposing raw patient data [33].

### 3.5. Federated Training and Supervision-Aware Aggregation

FedSAM3D follows a standard server-client optimization loop: at communication round $r$, the server broadcasts the current global parameters to participating clients; each

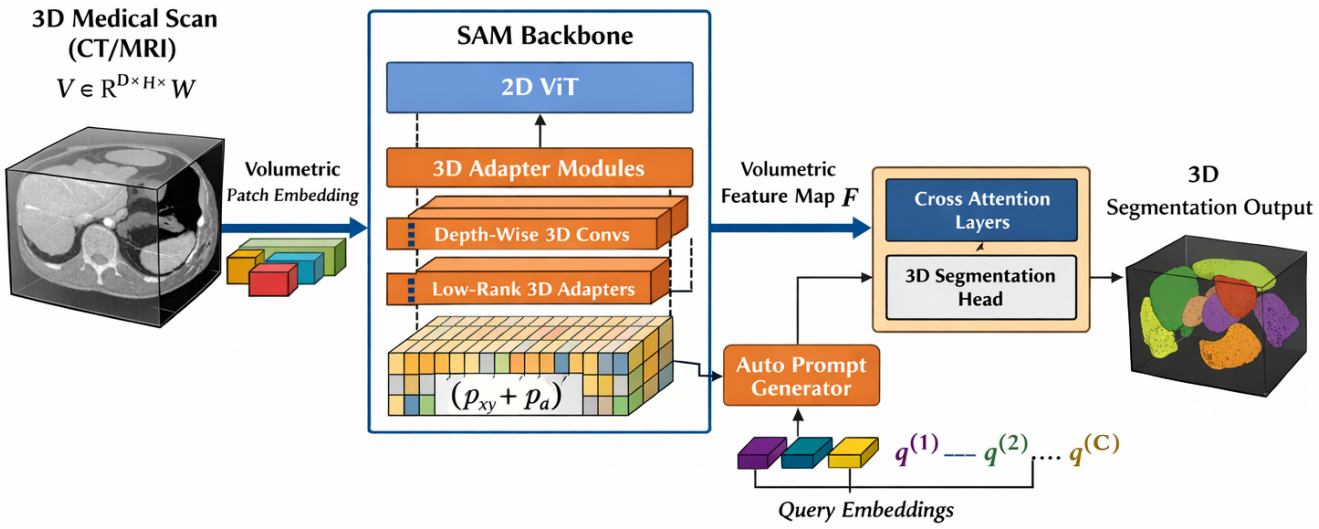


**Figure 5:** SAM backbone and parameter-efficient 3D adaptation in FedSAM3D. A volumetric scan $\mathbf{V} \in \mathbb{R}^{D \times H \times W}$ is tokenized via volumetric patch embedding and processed by a SAM image encoder initialized from a 2D ViT. To incorporate through-plane context with minimal communication overhead, FedSAM3D injects lightweight 3D adaptation modules (e.g., depth-wise 3D convolutions and low-rank 3D adapters) together with a factorized 3D positional encoding ($\mathbf{p}_{xy} + \mathbf{p}_d$), while keeping the majority of pretrained parameters frozen. The adapted encoder produces a volumetric feature map $\mathbf{F}$, which is fused with automatically generated class-conditioned query embeddings $\{\mathbf{q}^{(c)}\}_{c=1}^{C}$ via an auto-prompt generator and cross-attention layers in the 3D mask decoder/segmentation head to yield multi-organ 3D segmentation outputs.

client performs local training for a fixed budget (epochs or steps) and returns an update; the server aggregates client updates into a new global model. Unlike conventional federated segmentation where the full network may be updated, FedSAM3D updates only a parameter-efficient subset (volumetric adapters, prompt-related modules, and selected normalization parameters), while keeping the majority of the pretrained backbone frozen. This reduces the communication payload and mitigates client drift under non-IID data.

A key distinction in our setting is that annotation coverage varies across clients: single-organ cohorts provide dense supervision for only one organ, whereas multi-organ cohorts provide dense supervision for multiple organs. Consequently, dataset-size weighting (FedAvg) can overemphasize large single-organ clients even though they contribute no labeled signal for most organs. To better align aggregation with the effective supervised signal, we weight each client update by an "effective supervision mass" proportional to the amount of labeled training evidence it provides in the round (intuitively: number of supervised voxels across labeled organs under the patch-based training schedule). Let $\alpha_k^{(r)}$ denote the normalized aggregation weight for the client $k$ in round $r$. The server update is then:

$$\boldsymbol{\psi}^{(r+1)} = \sum_{k \in \mathcal{K}^{(r)}} \alpha_k^{(r)} \boldsymbol{\psi}_k^{(r)}, \tag{10}$$

where $\boldsymbol{\psi}_k^{(r)}$ denotes the locally updated parameters (for the trainable subset) returned by client $k$ and $\mathcal{K}^{(r)}$ is the set of participating clients. In practice, $\alpha_k^{(r)}$ is computed from each client's labeled coverage and local optimization budget, which improves stability when label density and organ coverage vary substantially across sites.

### 3.6. SAM Backbone and 3D Adaptation

Segment Anything Model (SAM) [22] is a prompt-conditioned segmentation paradigm that has demonstrated strong transferability across diverse 2D visual domains. SAM is composed of three interacting modules: (i) a visual feature extractor, (ii) a prompt representation module, and (iii) a mask inference head. The visual feature extractor is instantiated with a vision transformer (ViT) [7], which maps an input image to a high-dimensional latent feature grid. The prompt module converts user guidance (e.g., points, boxes, or sparse cues) into compact embeddings by combining positional codes with prompt-type-specific representations. The mask head then couples image features with prompt embeddings through self-attention and bidirectional cross-attention and produces segmentation logits via lightweight projection layers.

Despite its effectiveness on 2D natural-image benchmarks, directly applying SAM to volumetric medical imaging is non-trivial. First, a slice-wise 2D application ignores through-plane dependencies and can yield discontinuous predictions across adjacent slices, particularly for thin or low-contrast structures. Second, the modality and appearance shift between natural images and CT/MRI introduces a pronounced domain mismatch that degrades zero-shot generalization. Third, medical segmentation frequently requires consistent multi-organ delineation under limited

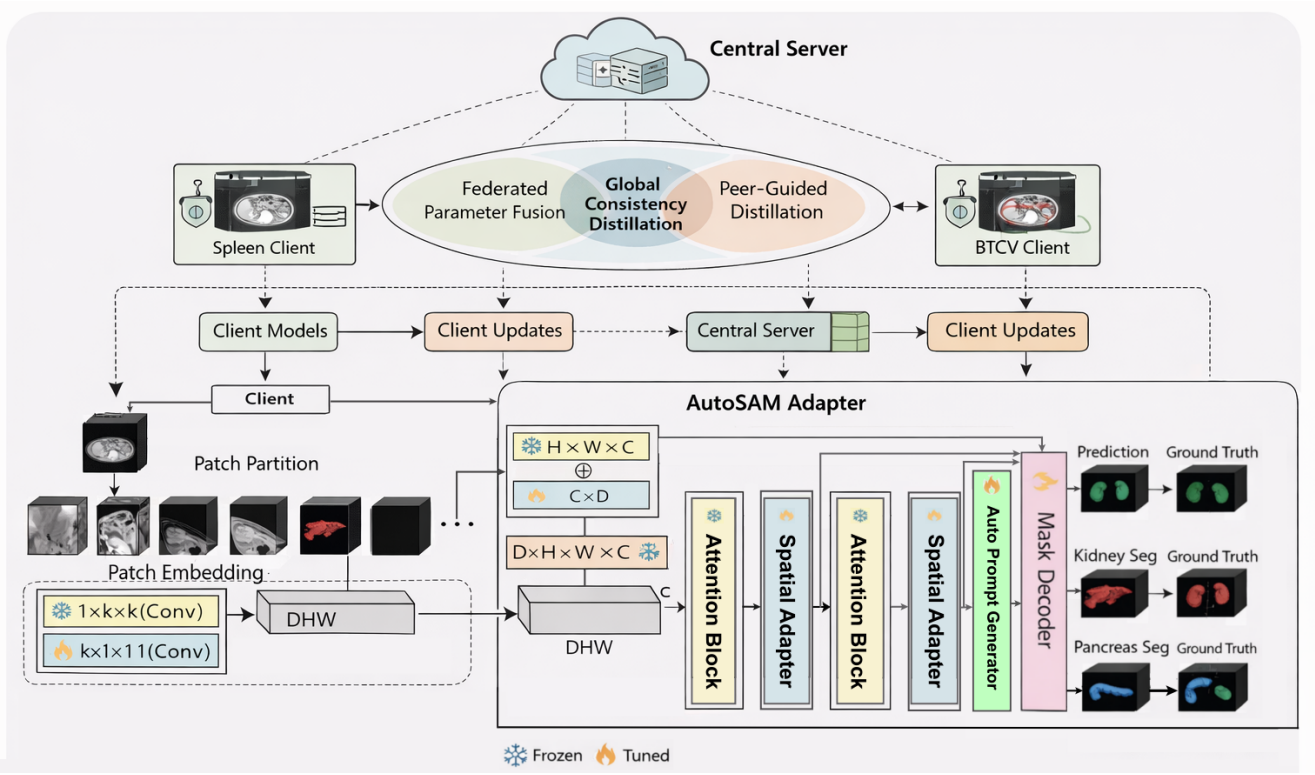


**Figure 6:** Architecture of FedSAM3D. A central server orchestrates cross-silo training by combining (i) federated parameter fusion of client updates with (ii) global consistency distillation and (iii) peer-guided distillation to transfer knowledge across heterogeneous, partially labeled clients (e.g., single-organ spleen and multi-organ BTCV cohorts). Each client processes a 3D volume via patch partitioning and a parameter-efficient volumetric patch embedding that factorizes in-plane and depth mixing, producing volumetric tokens of size $D \times H \times W \times C$. The proposed AutoSAM Adapter injects lightweight trainable components (spatial/adapter blocks and attention) into an otherwise frozen SAM-style backbone (snowflake vs. flame denotes frozen vs. tuned parameters) and uses an Auto Prompt Generator to synthesize class-conditioned queries for fully automatic multi-organ inference. A prompt-conditioned mask decoder outputs organ-wise 3D segmentations, supervised locally where labels exist and regularized via the coupled global-peer distillation pathways during federated optimization.

supervision, where purely prompt-driven 2D reasoning is insufficient to capture global 3D context.

To make SAM suitable for 3D multi-organ segmentation, FedSAM3D adopts a volumetric adaptation strategy that augments the SAM backbone with lightweight 3D-aware tuning modules and automatic prompt generation. Specifically, we treat a 3D scan $\mathbf{V} \in \mathbb{R}^{D \times H \times W}$ as a volumetric input and inject trainable, parameter-efficient blocks (e.g., depth-wise 3D convolutions or low-rank adapters) into selected layers of the feature extractor and/or mask head to model inter-slice correlations without retraining the full foundation model. In addition, the prompt stream is produced automatically from data-driven cues, enabling consistent training in fully automated pipelines where manual prompts are unavailable. These modifications allow FedSAM3D to retain SAM's prompt-conditioned decoding advantages while learning volumetric continuity and domain-specific representations required for reliable medical segmentation under heterogeneous, partially labeled federated data.

### 3.7. Volumetric Input Handling and 3D Adaptation

FedSAM3D extends the prompt-conditioned SAM backbone to operate on volumetric CT/MRI scans while preserving the representational benefits of large-scale 2D pre-training. A direct 2D deployment on a 3D scan is suboptimal because it (i) treats each slice independently, weakening through-plane coherence, and (ii) inherits a 2D-centric inductive bias that does not explicitly model volumetric neighborhood structure. Furthermore, naively converting all operators to dense 3D variants substantially increases memory consumption and communication cost, which is undesirable in federated settings. Our design therefore follows two guiding principles: (a) introduce explicit 3D spatial awareness, and (b) retain compatibility with pre-trained 2D parameters by restricting training to a small set of incremental modules, as shown in Figure 6.

In a 2D ViT, positional codes are commonly defined over the in-plane token grid. For a volume $\mathbf{V} \in \mathbb{R}^{D \times H \times W}$,

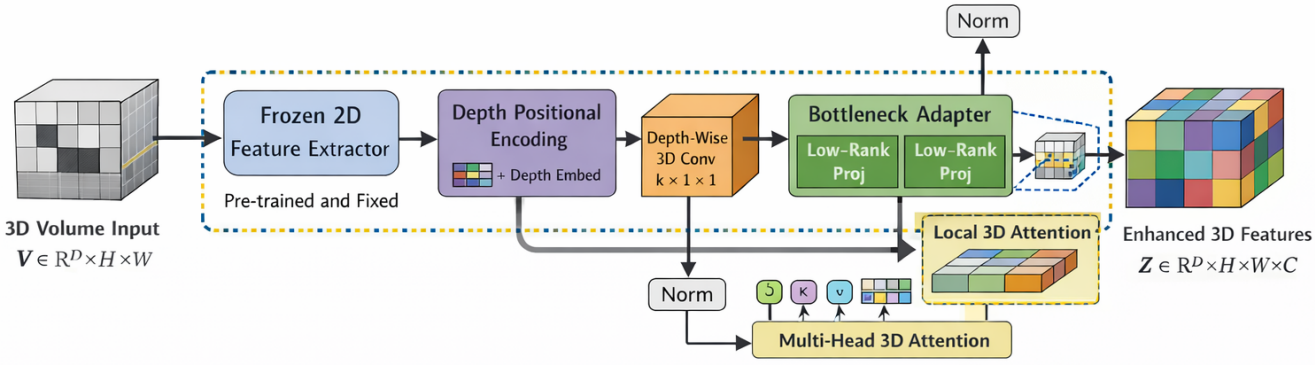


**Figure 7:** Spatial adapter in FedSAM3D. A 3D volume $\mathbf{V} \in \mathbb{R}^{D\times H\times W}$ is encoded by a frozen 2D feature extractor, augmented with learnable depth positional embeddings, and enhanced via a depth-wise $k\times1\times1$ 3D convolution and a low-rank bottleneck adapter. In parallel, normalized multi-head local 3D attention performs volumetric token mixing. The module outputs enhanced features, $\mathbf{Z} \in \mathbb{R}^{D\times H\times W\times C}$ using a parameter-efficient trainable model for federated communication.

we augment the in-plane positional code with an additional depth-index embedding so that each token at a location $(d, h, w)$ receives a separable 3D positional signal, $\mathbf{p}(d, h, w) = \mathbf{p}_{\text{xy}}(h, w)+\mathbf{p}_{\text{d}}(d)$, where $\mathbf{p}_{\text{xy}}$ is initialized from the pre-trained 2D table (kept frozen for continuity) and $\mathbf{p}_{\text{d}}$ is a trainable lookup table initialized to zeros. This factorized design introduces a through-plane identity with minimal additional parameters and avoids disrupting the original 2D token geometry.

To embed volumetric patches efficiently, we replace a dense $k \times k \times k$ projection with a factorized operator that separates in-plane aggregation from depth mixing. Concretely, we apply a frozen in-plane projection $(1 \times k \times k)$ initialized from the 2D patch embedder, followed by a lightweight depth mixer $(k \times 1 \times 1)$ implemented as a depth-wise 3D convolution to limit trainable parameters, $\mathbf{z} = \text{DWConv}_{k\times1\times1}\Big(\text{Conv}_{1\times k\times k}(\mathbf{V})\Big)$.

This approximation preserves the effect of a volumetric receptive field while keeping the computational and communication overhead tractable. Self-attention over all $DHW$ tokens is memory-intensive. We therefore reshape the token sequence from a 2D layout $[B, HW, C]$ to a volumetric layout $[B, DHW, C]$ and employ windowed attention to bound complexity as $\text{Attn}_{3\text{D}} : [B, DHW, C] \to [B, DHW, C]$, where attention is computed within local 3D windows and shifted between layers to allow cross-window interaction, following the sliding/shifted-window principle used in SwinUNETR [40]. This modification enables volumetric context aggregation while preventing quadratic scaling in token count. Replacing all 2D operators with 3D counterparts can lead to overfitting on small medical datasets and increase memory/communication costs. Instead, we introduce a compact bottleneck adapter inserted into selected transformer blocks as shown in Figure 7. For an intermediate token feature matrix $\mathbf{U} \in \mathbb{R}^{N\times C}$ (with $N$ tokens and $C$ channels), the adapter is defined as $\text{VA}(\mathbf{U}) = \mathbf{U}+\phi\big(\mathbf{U}\mathbf{A}_{\downarrow}\big)\,\mathbf{A}_{\uparrow}$, where $\mathbf{A}_{\downarrow} \in \mathbb{R}^{C\times r}$ and $\mathbf{A}_{\uparrow} \in \mathbb{R}^{r\times C}$ are trainable projection matrices with rank $r \ll C$, and $\phi(\cdot)$ is a nonlinearity (e.g., GELU/ReLU).

To improve spatial sensitivity, we further insert a lightweight depth-wise 3D convolution in the reduced space as:

$$\text{VA}(\mathbf{U}) = \mathbf{U} + \mathbf{A}_{\uparrow}\,\text{DWConv}_{3\text{D}}\Big(\phi\big(\mathbf{U}\mathbf{A}_{\downarrow}\big)\Big), \tag{11}$$

which allows the adapter to model local 3D neighborhoods while keeping the main pre-trained backbone unchanged. During federated training, only the incremental 3D components (depth positional table $\mathbf{p}_{\text{d}}$, depth mixer, windowed-attention configuration, volumetric adapters in (11), and normalization layers) are updated, while the majority of pre-trained SAM parameters remain frozen. This selective tuning reduces GPU memory demand, lowers communication payloads, and improves robustness under partial labels by avoiding large-scale over-parameterization, while still enabling the model to learn volumetric continuity and modality-specific features required for accurate multi-organ delineation [22].

### 3.8. Auto Prompt Generator

SAM-style prompt-conditioned decoders rely on a shared positional reference: prompt and image tokens corresponding to the same spatial location are encoded with compatible positional codes, and cross-attention is then used to align sparse guidance with dense visual features. While this design is effective for 2D interactive segmentation, two issues arise in volumetric multi-organ medical settings. First, the token count grows substantially in 3D, and repeated cross-attention between dense volumetric tokens and multiple prompts can increase compute and promote feature over-smoothing, yielding less discriminative representations and flatter confidence maps. Second, prompt-driven pipelines are operationally constrained: (i) generating prompts for many organs is time-consuming and scales poorly with the number of targets, and (ii) segmentation quality becomes sensitive to prompt precision, which often requires domain expertise that is not consistently available in practice. These

limitations reduce the usability and scalability of prompt-based segmentation for routine multi-class clinical applications. To overcome these constraints, FedSAM3D introduces an Auto Prompt Generator (APG), which replaces manual prompt specification with data-driven query embeddings derived directly from volumetric feature maps. Rather than relying on static positional prompt encodings, APG learns to synthesize a compact set of class-aware prompt tokens from the adapted 3D backbone's output (as shown in Figure 6), enabling fully automatic training and inference. Concretely, given the volumetric feature tensor $\mathbf{F}_{k,n}$ produced after the final attention and volumetric tuning blocks, APG predicts a set of prompt vectors $\{\mathbf{q}_{k,n}^{(c)}\}_{c=1}^{C}$ that are subsequently fed into the SAM mask inference head to decode organ masks.

APG is implemented as a compact fully convolutional network, inspired by the efficient 3D encoder-decoder design of 3D U-Net [37]. It employs parameter-light 3D convolutions to preserve volumetric context while maintaining low computational overhead. Formally, APG is a mapping as:

$$\{\mathbf{q}_{k,n}^{(c)}\}_{c=1}^{C} = g_{\xi}\left(\mathbf{F}_{k,n}\right), \tag{12}$$

where $g_{\xi}(\cdot)$ denotes the APG with parameters $\xi$. Each $\mathbf{q}_{k,n}^{(c)} \in \mathbb{R}^{d_q}$ is a class-conditioned prompt token (or a small set of tokens per class) that captures spatially aware and context-specific cues for organ $c$. To reduce the risk of prompt collapse (all prompts becoming similar) and to improve calibration under partial labels, APG is trained jointly with the segmentation objective using: (i) class-balanced sampling of target queries, (ii) confidence gating when APG-derived prompts are used for unlabeled organs, and (iii) lightweight regularization (e.g., norm or diversity penalties) to encourage prompt diversity across organs. These design choices address practical shortcomings of naive automatic prompting: unstable multi-class prompt learning in 3D and sensitivity to ambiguous features.

## 4. Experiments and Results

### 4.1. Dataset

#### *4.1.1. Spleen Dataset*

We use the Spleen [31] dataset from the Medical Segmentation Decathlon (MSD) as a representative single-organ abdominal CT cohort in our partial-label federated setting. The dataset contains 61 contrast-enhanced abdominal CT volumes with voxel-wise delineations provided only for the spleen. Following the official MSD split, 41 labeled cases are used for model development, while 20 additional cases are reserved for benchmark-style testing (labels are not publicly distributed in the standard release). In this work, we train and tune hyperparameters using only the labeled subset by constructing a patient-wise internal validation split from the 41 training volumes. The acquisitions exhibit noticeable variability in scan geometry and resolution. Reported voxel spacing typically ranges from approximately $[0.5 \times 0.5 \times 1.0]$ mm to $[1.0 \times 1.0 \times 2.0]$ mm, and volumes commonly have in-plane size $512 \times 512$ with about 80–200 axial slices depending on field-of-view and protocol. Although other abdominal structures (e.g., liver, kidneys, pancreas) are visible in many scans, annotations are restricted to the spleen. This label sparsity aligns with our study setting, where each client provides supervision for a strict subset of the global organ vocabulary, and all non-annotated organs are treated as unlabeled (not background) during training, consistent with the masked supervision design in FedSAM3D.

#### *4.1.2. Liver Dataset*

We use the Liver [2] dataset from the medical segmentation which is derived from the Liver Tumor Segmentation (LiTS) data, as a liver-focused client in our cross-dataset federation. The dataset contains 201 contrast-enhanced abdominal CT scans acquired across multiple imaging centers, leading to variability in acquisition protocols and scanner settings. The original annotations include voxel-wise labels for the liver parenchyma and associated lesions. However, to remain consistent with our partial-label federation design (where each site supervises only a subset of the global organ vocabulary), we retain only the liver organ mask for supervised training and validation in this client. The official benchmark split provides 131 labeled volumes for training and 70 volumes for testing. The scans exhibit heterogeneous spatial resolution, with voxel spacing commonly reported in the range of approximately $[0.6 \times 0.6 \times 1.0]$ mm to $[1.5 \times 1.5 \times 2.5]$ mm and typical in-plane dimensions of $512 \times 512$ with about 100–300 axial slices depending on field-of-view [3; 4]. Importantly, other abdominal structures (e.g., spleen, kidneys, pancreas) are frequently visible in the same volumes but are not annotated in this task.

#### *4.1.3. Pancreas Dataset*

We use the Pancreas Segmentation [2] dataset that representative single-organ abdominal CT cohort in our partial-label federated setting. The dataset contains 420 contrast-enhanced abdominal CT volumes collected across multiple institutions, resulting in heterogeneous acquisition protocols and scanner configurations. Voxel-wise pancreas annotations were produced by expert readers and are used as the reference standard for supervised training and evaluation. The official release provides a predefined split with 282 training cases and 139 test cases. Since ground-truth masks for the test partition are not always publicly distributed in standard benchmarking pipelines, we develop the model using the labeled training subset and construct an internal patient-wise validation split from it. The scans exhibit variable voxel spacing, with typical resolutions spanning approximately $[0.7 \times 0.7 \times 1.5]$ mm to $[1.25 \times 1.25 \times 2.5]$ mm, and volume sizes commonly around $512 \times 512 \times [120–300]$ voxels. Although multiple abdominal structures are visible in the field-of-view, supervision is provided only for the pancreas in this task. Therefore, when the dataset is instantiated as a federated client, its labeled set is restricted to $\mathcal{O}_k = \{\text{pancreas}\}$, and all other organs are treated as unlabeled rather than background in the masked objective.

#### 4.1.4. Kidney Dataset

We use the Kidney [12] dataset as a kidney-focused abdominal CT cohort within our partial-label federated protocol. It comprises contrast-enhanced abdominal CT volumes acquired across multiple clinical centers, introducing variability in scanner vendors, acquisition settings, and reconstruction parameters. The dataset provides expert voxel-level annotations for the kidneys and renal tumors. In this work, we retain only the kidney masks and discard the tumor labels to align the client supervision with a single-organ partial-label setting (i.e., $\mathcal{O}_k = \{\text{kidney}\}$) and to avoid confounding effects from lesion-specific labeling policies.

Following the commonly used labeled split, we use 210 annotated cases for development and evaluation, with 168 for training and 42 reserved for testing. The scans show heterogeneous voxel spacing, typically ranging from approximately $[0.5 \times 0.5 \times 1.0]$ mm to $[1.25 \times 1.25 \times 2.0]$ mm, and volume sizes commonly around $512 \times 512 \times [80\text{–}250]$ voxels. While additional abdominal structures are present in the field of view, ground-truth annotations in this client are restricted to kidney tissue, which directly reflects the label-space fragmentation addressed by FedSAM3D. When instantiated as a client in the federation, this dataset contributes supervised gradients only for the kidney class, while all other organs in the global vocabulary are treated as unlabeled (not background) under the masked loss.

#### 4.1.5. BTCV Multi-Organ Dataset

We use the Beyond the Cranial Vault (BTCV) dataset [23], a contrast-enhanced abdominal CT benchmark with expert voxel-wise annotations for 13 anatomical structures (e.g., liver, spleen, pancreas, left/right kidneys, stomach, major vessels, gallbladder, esophagus, and adrenal glands). The cohort exhibits substantial acquisition heterogeneity across sites (scanner vendor, reconstruction, and field-of-view), making it suitable for evaluating multi-organ segmentation robustness under realistic domain shift. In our federation, BTCV serves as the sole client with dense multi-organ supervision, while the remaining clients provide partial labels. This configuration directly probes whether the proposed masked supervision and coupled global-peer distillation prevent "missing class = background" failure modes and preserve multi-organ competence during aggregation across mismatched label spaces. Following common practice, we use the labeled subset and evaluate on 30 annotated volumes with patient-wise splits (e.g., 18/12 train/test), reporting metrics only for organs with available ground truth and emphasizing overlapping abdominal targets (liver, spleen, kidneys, pancreas).

### 4.2. Data Preprocessing and Augmentation

To reduce avoidable feature-shift heterogeneity across clients and to make volumetric prompting stable, all sites apply the same deterministic preprocessing pipeline before local training. Augmentations are applied oon the fly during local optimization and are not shared across sites. This design keeps the raw data private while improving robustness of the 3D-adapted SAM backbone and the Auto Prompt Generator (APG) under non-IID imaging. Each CT volume is first reoriented to a canonical axis convention and resampled to a fixed isotropic spacing to ensure consistent voxel geometry across datasets and institutions. We use a single target spacing for all clients to keep the volumetric tokenization and windowed attention configuration consistent. In practice, isotropic resampling in the range of 1.0–1.5 mm is commonly used for abdominal CT segmentation, depending on memory constraints. Following recent large-scale 3D CT segmentation practice, we additionally adopt patch-based training with fixed-size sub-volumes (e.g., $128^3$ voxels) to control GPU memory and to align with our SAM-style volumetric tokenization. For each client, we crop to the body region (a foreground crop) and pad as needed to support the extraction of fixed-size patches. When a structure is small (e.g., a pancreas or a lesion), we use class-aware sampling to increase the likelihood that sampled patches contain foreground voxels, thereby stabilizing APG prompt learning and reducing empty-patch updates.

To mitigate scanner/protocol differences, intensities are clipped to suppress extreme outliers and then normalized per volume. Specifically, we apply a CT intensity-range scaling (linear mapping to a bounded range, followed by standardisation) and maintain the same normalisation scheme across clients to avoid introducing client-specific scaling artefacts that can amplify drift during federated fusion. This normalization is applied before feeding volumes into the 3D-adapted SAM encoder so that prompt-conditioned decoding operates on comparable feature statistics across sites. All datasets are mapped into the global organ vocabulary $\mathcal{O}$ used by FedSAM3D. For client $k$, only the locally available labels $\mathcal{O}_k$ are retained as supervised targets; voxels belonging to organs outside $\mathcal{O}_k$ are treated as unlabeled rather than background. Implementation-wise, we store a per-class availability mask $\mathbb{I}_k(c)$ and compute supervised losses only for $c \in \mathcal{O}_k$ as in Equation (**??**). This prevents false-negative gradients for missing organs and keeps APG training consistent with the masked objective.

During local training, we apply stochastic 3D geometric transformations to improve generalisation across site variability. Augmentations include random spatial cropping, random flips, and mild rotations/scaling, applied identically to the volume and the corresponding label masks. These transforms are compatible with patch-based training and help the APG learn prompts that are less sensitive to small spatial shifts.To improve robustness to contrast and noise differences across scanners, we apply intensity perturbations such as random intensity scaling/shift and additive noise. Intensity transforms are restricted to preserve anatomical plausibility while providing sufficient diversity for non-IID sites.FedSAM3D performs fully automatic prompting: APG generates class-conditioned query prompts from the augmented volumetric features, so prompts remain consistent with all applied transforms. When auxiliary prompt cues are used for debugging or ablations (e.g., mask-derived boxes/points), they are derived after augmentation so that the prompt encoder and mask decoder receive geometrically

aligned guidance. All preprocessing steps are deterministic and identical across clients, while augmentations are randomized per mini-batch and remain local to each client. This separation improves federated stability by reducing avoidable preprocessing-induced domain gaps while preserving augmentation diversity across sites.

### 4.3. Evaluation Metrics

We evaluate FedSAM3D on all datasets using voxel overlap and surface distance metrics that are standard in volumetric medical segmentation. In our partial-label federation, each client $k$ has ground-truth annotations only for its local label subset $\mathcal{O}_k$ (e.g., spleen, kidney, and multi-organ labels for BTCV). Therefore, metrics are computed only for classes with available ground truth and are excluded for missing classes to avoid conflating "unlabeled" with "negative" labels. For each metric, we first compute per-volume and per-class scores, then summarize them as mean±std across test volumes for each client, and finally report a macro-average across clients (equal weight per client) to avoid dominance by larger datasets. This reporting protocol is consistent with recent 2025 segmentation benchmarks and federated segmentation studies that use Dice and robust surface distances for volumetric evaluation [38; 14].

For a client $k$, test volume $n$, and class $c \in \mathcal{O}_k$, let $\widehat{\mathbf{Y}}_{k,n}^{(c)} \in \{0,1\}^{D\times H\times W}$ be the binarized prediction (thresholded from $\widehat{\mathbf{P}}_{k,n}^{(c)}$) and $\mathbf{G}_{k,n}^{(c)}$ be the ground truth. The DSC is

$$\mathrm{DSC}\Big(\widehat{\mathbf{Y}}_{k,n}^{(c)}, \mathbf{G}_{k,n}^{(c)}\Big) = \frac{2\,\Big|\widehat{\mathbf{Y}}_{k,n}^{(c)} \cap \mathbf{G}_{k,n}^{(c)}\Big|}{\Big|\widehat{\mathbf{Y}}_{k,n}^{(c)}\Big| + \Big|\mathbf{G}_{k,n}^{(c)}\Big| + \epsilon}, \tag{13}$$

where $|\cdot|$ denotes the number of foreground voxels and $\epsilon$ prevents division by zero [6].

Overlap metrics can mask boundary errors, especially for small organs and lesions (e.g., pancreas and tumours). We therefore report the 95th percentile Hausdorff distance (HD95) in millimeters. Let $\mathcal{S}(\widehat{\mathbf{Y}})$ and $\mathcal{S}(\mathbf{G})$ be the sets of surface points extracted from the predicted and ground-truth masks (after mapping to physical coordinates using voxel spacing). For a surface point $x$, define its distance to a surface set $\mathcal{S}$ as $d(x,\mathcal{S}) = \min_{y\in\mathcal{S}} \|x-y\|_2$. The directed surface distance set from prediction to ground truth as $\mathcal{D}_{\widehat{\mathbf{Y}}\to\mathbf{G}} = \{d(x,\mathcal{S}(\mathbf{G})) \mid x \in \mathcal{S}(\widehat{\mathbf{Y}})\}$, and analogously $\mathcal{D}_{\mathbf{G}\to\widehat{\mathbf{Y}}}$. HD95 is then defined as

$$\mathrm{HD95}(\widehat{\mathbf{Y}}, \mathbf{G}) = \max\Big(\mathrm{perc}_{95}\big(\mathcal{D}_{\widehat{\mathbf{Y}}\to\mathbf{G}}\big),\ \mathrm{perc}_{95}\big(\mathcal{D}_{\mathbf{G}\to\widehat{\mathbf{Y}}}\big)\Big), \tag{14}$$

where $\mathrm{perc}_{95}(\cdot)$ denotes the 95th percentile. Using a percentile (rather than the maximum) reduces sensitivity to isolated outliers and is widely used for 3D medical segmentation evaluation [8; 39].

Let $\mathcal{T}_k$ be the set of labeled test volumes on client $k$ and $\mathcal{O}_k$ its labeled class set. For a metric $m(\cdot)$ (DSC or HD95), we compute the per-client macro score

$$\overline{m}_k = \frac{1}{|\mathcal{O}_k|}\sum_{c\in\mathcal{O}_k}\left(\frac{1}{|\mathcal{T}_k|}\sum_{n\in\mathcal{T}_k} m\Big(\widehat{\mathbf{Y}}_{k,n}^{(c)}, \mathbf{G}_{k,n}^{(c)}\Big)\right), \tag{15}$$

and report mean±std across $n \in \mathcal{T}_k$ for each class when organ-wise results are shown. The federation-level macro score is

$$\overline{m} = \frac{1}{K}\sum_{k=1}^{K}\overline{m}_k, \tag{16}$$

which gives equal weight to each client and aligns with the goal of robust cross-site performance rather than dataset-size dominance.

In addition to segmentation accuracy, we report two efficiency indicators that are specific to FedSAM3D's parameter-efficient design:

- **Communication payload (MB/round).** For each round, we measure the number of bytes uploaded from a client to the server for trainable parameters (and, when enabled, lightweight class experts) and convert it to MB. This reflects the benefit of updating only the volumetric adapters/APG modules rather than the full SAM backbone.
- **Peak GPU memory (GB).** We record the maximum GPU memory allocated during local training. This metric is relevant for volumetric prompting because memory is a limiting factor for 3D attention and patch-based training.

These indicators are reported alongside DSC/HD95 to characterize the accuracy-efficiency trade-off in federated deployment [38].

### 4.4. Implementation Details

All models are implemented in PyTorch with mixed-precision training (AMP) [36] and medical-imaging preprocessing/augmentation using MONAI transforms [5]. Unless otherwise stated, training is performed on a single GPU (24 GB memory class) with CUDA-enabled acceleration; multi-GPU execution is not required because FedSAM3D updates only lightweight adaptation modules and uses patch-based 3D training. We instantiate a cross-dataset federation with $K = 5$ clients, where each dataset corresponds to one privacy-isolated site: Spleen, Liver±Tumor, BTCV (13 organs), Pancreas±Tumor, and Kidney±Tumor. Each client optimizes the masked objective in Equation (**??**) over its local label set $\mathcal{O}_k$ while excluding missing classes from supervised loss via $\mathbb{I}_k(c)$. For each dataset, we use the official labeled training subset and construct an internal validation split from the labeled training cases (patient-wise) to tune hyperparameters and early stopping; BTCV follows the common 18/12 train/test split, while the remaining datasets use an 80/20 split of labeled training cases for train/validation when official test labels are not publicly available.

FedSAM3D initializes the SAM image encoder and mask decoder from publicly released SAM weights [22]. To control memory and communication, we freeze the majority of SAM parameters and update only: (i) the proposed volumetric adaptation modules (depth positional table $\mathbf{p}_{\mathrm{d}}$, depth mixer, and volumetric adapters), (ii) normalization layers (client updates are applied to affine parameters; running statistics are kept local when batch normalization is used, following the FedBN principle [25]), and (iii) the Auto Prompt Generator (APG). This yields a parameter-efficient trainable set typically below a few percent of the full backbone, which substantially reduces per-round payload.

Adapters follow Equation (11) with bottleneck rank $r \in \{8, 16\}$ (default $r = 16$), and the depth-wise 3D convolution uses kernel size $3 \times 3 \times 3$ in the reduced space. Adapters are inserted into the last $L_a$ transformer blocks of the SAM encoder (default $L_a = 6$) to bias adaptation toward higher-level semantics while retaining low-level filters from pretraining. Windowed attention is used for volumetric token sequences to bound memory growth, conceptually aligned with shifted-window volumetric transformers [11].

## 4.5. Training Perimeter

We instantiate FedSAM3D on a cross-dataset federation with $K = 5$ clients, each corresponding to one dataset/site: (i) Spleen, (ii) Liver and Liver-Tumour, (iii) BTCV 13 abdominal organs, (iv) Pancreas and Pancreas-Tumour, and (v) Kidney and Kidney-Tumour. The global label vocabulary $\mathcal{O} = \{1, \dots, C\}$ is defined as the union of all annotated structures across the five sites, while each client optimizes only on its locally available subset $\mathcal{O}_k \subseteq \mathcal{O}$ using the supervision mask $\mathbb{I}_k(c)$ (Equation (**??**)). This protocol directly matches the partial-label regime studied in recent federated segmentation work [15; 43; 19; 38].

Training proceeds for $R$ communication rounds. Unless otherwise stated, we use full participation ($\mathcal{K}^{(r)} = \{1, \dots, K\}$) to focus on label-space mismatch and non-IID appearance differences across datasets. For robustness analysis, we also report partial participation, where $|\mathcal{K}^{(r)}| < K$ (randomly sampled each round), following common FL evaluation practice [10]. At each round $r$, the server broadcasts the current global parameters $\boldsymbol{\psi}^{(r)}$ to all participating clients. Dataset sizes differ substantially. To avoid large sites dominating the update frequency, each client performs a fixed local optimization budget per round: $U$ gradient steps on randomly sampled 3D patches. Let $\boldsymbol{\psi}_k^{(r,t)}$ denote the client parameters at local step $t \in \{0, \dots, U\}$.

FedSAM3D updates only the parameter-efficient subset (volumetric adapters, depth positional codes, APG, and selected normalization parameters), while keeping the majority of the SAM backbone frozen [22]. Each client transmits either the tuned weights or a delta update which reduces communication compared to full-model exchange and aligns with parameter-efficient federated segmentation designs [38; 29]. When peer-guided distillation is enabled, clients additionally upload lightweight class experts $\{\boldsymbol{\omega}_k^{(c)}\}_{c \in \mathcal{O}_k}$.

## 4.6. Experimental Setup

We evaluate FedSAM3D in a cross-dataset, cross-silo federation of heterogeneous abdominal CT benchmarks with fragmented label spaces. The federation comprises $K = 5$ clients: spleen, liver±tumor, BTCV (13 organs), pancreas ±tumor, and kidney±tumor. All methods use the same deterministic preprocessing and on-the-fly augmentation and are evaluated using DSC/HD95 and efficiency indicators (payload per round and peak GPU memory). Training follows the parameter-efficient communication protocol and coupled distillation objectives. The global label vocabulary $\mathcal{O}$ is the union of all annotated structures, while client $k$ provides supervision only for $\mathcal{O}_k \subseteq \mathcal{O}$ and supervised losses are computed only for available classes via $\mathbb{I}_k(c)$.

For each dataset, we use the official labeled training set with a patient-wise internal validation split for hyperparameter selection and early stopping. BTCV follows the standard 18/12 train/test split of the labeled subset; for MSD tasks without public test labels, results are reported on a held-out internal test split. Federation-level performance is obtained by evaluating the global model on each client's test set for classes in $\mathcal{O}_k$ only and macro-averaging across clients (Equation (16)). Unless otherwise stated, we use full participation (all clients per round). Robustness is assessed under partial participation rates $\rho \in \{0.5, 0.25\}$, with $|\mathcal{K}^{(r)}| = \lceil \rho K \rceil$ clients sampled uniformly per round.

We compare against: (i) Local-only training per client (masked objective on $\mathcal{O}_k$), (ii) a centralized pooled-data upper bound, (iii) federated optimization baselines (FedAvg, FedProx, SCAFFOLD, and a FedBN-style variant), and (iv) partial-label federated segmentation baselines (e.g., UFPS, FedIA, and Fed-MENU). For fair comparison, all methods share identical data handling, patching/inference settings, communication rounds $R$, and per-round local optimization budgets. Hyperparameters are selected using only client-local validation data; global FL hyperparameters (e.g., learning rate, weight decay, $\mu$, $\beta$, $\gamma$) are tuned by maximizing mean validation DSC across clients/classes, while FedSAM3D uses fixed $(\tau, \tau_{\mathrm{P}})$ and $m_k$ across datasets. Unless otherwise stated, results are reported as mean±std over three independent runs with fixed random seeds.

## 4.7. Comparative Models

To evaluate the performance of FedSAM3D, we compare against SoTA models such as FednnU-Net [38], UFPS [15], FedIA [43], KD-Synth [19] and FedFMS [29] that align with our setting: federated optimization for volumetric segmentation under cross-silo heterogeneity, heterogeneous and partial supervision across clients, distillation-based knowledge transfer, and parameter-efficient federated adaptation of large backbones. Unless a baseline requires a different protocol, all methods are evaluated under the same federation configuration ($K = 5$ clients), communication budget ($R$ rounds), fixed per-round local optimization budget, identical preprocessing/augmentation, patch-based training, and sliding-window inference.

**Table 1**
Evaluate the partial-label correctness on BTCV dataset. DSC and HD95 are computed only for labeled organs at each client; BTCV is macro-averaged across 13 organs. Mean±std over multiple random seeds.

| Supervision strategy | DSC ↑ (mean±std) / HD95 ↓ (mm) | | | | |
|---|---|---|---|---|---|
| | Spleen | Liver | Pancreas | Kidney | BTCV (13-org macro) |
| Naïve multi-class loss | 0.941±0.022 / 4.1 | 0.955±0.019 / 5.0 | 0.804±0.062 / 13.1 | 0.926±0.028 / 6.2 | 0.771±0.037 / 10.4 |
| Ignore-unlabeled voxels | 0.952±0.016 / 3.4 | 0.966±0.013 / 4.1 | 0.837±0.047 / 10.1 | 0.941±0.019 / 4.8 | 0.818±0.025 / 8.1 |
| **Masked supervision (ours)** | **0.958±0.014 / 3.0** | **0.971±0.011 / 3.7** | **0.855±0.043 / 9.2** | **0.953±0.016 / 4.0** | **0.842±0.022 / 6.9** |

We include a strong system-level federated baseline by adopting federated nnU-Net (FednnU-Net) [38], which extends the nnU-Net-style self-configuration and engineering to cross-institution training and serves as a competitive reference for practical 3D medical segmentation in FL. To specifically evaluate robustness under fragmented supervision and label-space mismatch, we compare against UFPS [15], which formalizes partially supervised federated segmentation and introduces unified label learning with stabilization mechanisms tailored to incomplete annotations, and FedIA [43], which models heterogeneous annotation completeness and incorporates completeness-aware strategies to mitigate the adverse effects of missing supervision during federated training. Since missing-class knowledge transfer is central to our setting, we further include a recent distillation-based FL baseline that distils client predictors into a global model using proxy data augmented with server-side synthesis to reduce proxy-domain mismatch, enabling communication-efficient training under heterogeneity (KD-Synth) [19]. Finally, to position FedSAM3D against foundation-model-based federated adaptation, we include FedFMS [29], which studies federated deployment of SAM-style models via parameter-efficient tuning modules to reduce communication and training cost; we instantiate this approach under the same volumetric patching and evaluation protocol to isolate the effect of FedSAM3D's volumetric prompting, partial-label masking, and coupled global-peer distillation.

## 4.8. Ablation Study

### *4.8.1. Ablation Study on Partial-Label Correctness*

Partial-label federated multi-organ segmentation is vulnerable to the missing-as-background assumption: single-organ clients generate systematic false-negative gradients for unannotated organs that are nevertheless present, inducing class collision during aggregation and suppressing rare/site-specific structures. To isolate this factor, all variants use the same FedSAM3D backbone (3D adapters + APG), preprocessing/augmentation, and federated schedule; only the supervised objective is changed. We compare: (i) Naïve multi-class loss, where absent organ masks are set to all-zero (missing treated as background); (ii) Ignore-unlabeled voxels, where supervision is restricted to labeled regions; and (iii) Masked supervision (ours), which applies class-availability masking $\mathbb{I}_k(c)$ and optimizes $\sum_c \mathbb{I}_k(c)\,\mathcal{E}(\widehat{P}^{(c)}, G^{(c)})$ (Equation (7)).

Table 1 shows that treating missing organs as background produces the largest degradation across all clients, with the most severe impact on BTCV (multi-organ) and on anatomically challenging structures such as the pancreas. Compared with masked supervision, the naïve variant reduces BTCV macro DSC by approximately 7.1 points (0.842→0.771) and increases HD95 by ∼3.5 mm (6.9→ 10.4 mm). This behavior is consistent with class collision: single-organ clients contribute gradients that systematically suppress logits for organs they do not annotate, biasing the aggregated global model toward under-segmentation for precisely those structures that are sparse or client-specific. The effect is amplified on BTCV because many organs lack supervision for most clients; thus, the federation receives contradictory signals, with the same anatomy implicitly labeled as background at some sites and as foreground at the BTCV site. Beyond mean performance, the naïve setting exhibits noticeably higher variance (e.g., pancreas DSC standard deviation ≈ 0.062), reflecting optimization instability under federated aggregation.

Figure 8 shows the concrete failure modes behind the quantitative gap in Table 1. When missing organs are treated as background (naïve multi-class), the model receives systematic false-negative gradients for anatomically present but unlabeled structures, which manifests as organ suppression (e.g., kidney) and fragmented, leaking predictions indicative of class collision. Ignoring unlabeled voxels reduces this suppression but removes dense one-vs-rest constraints outside labeled regions, weakening calibration and producing over-segmentation. In contrast, masked supervision applies class-availability masking $\mathbb{I}_k(c)$ to exclude missing organs from supervision while retaining dense supervision for annotated targets across the full voxel grid, yielding organ-aligned and spatially consistent segmentations.

Ignoring unlabeled voxels alleviates false-negative gradients for missing organs, yielding a clear improvement over the naïve baseline. However, it remains consistently inferior to $\mathbb{I}_k(c)$ masking (BTCV DSC: 0.818 vs 0.842; pancreas DSC: 0.837 vs 0.855). The gap is explained by a different failure mode: while the ignore strategy avoids penalizing missing-class predictions, it also discards a large fraction of negative/background evidence outside labeled regions. In single-organ cohorts, where the labeled foreground occupies a small volume fraction, this can weaken calibration and encourage over-segmentation or boundary leakage into adjacent tissues, which is reflected in higher HD95 (e.g., kidney: 4.8 mm vs 4.0 mm). In contrast, $\mathbb{I}_k(c)$ masking preserves

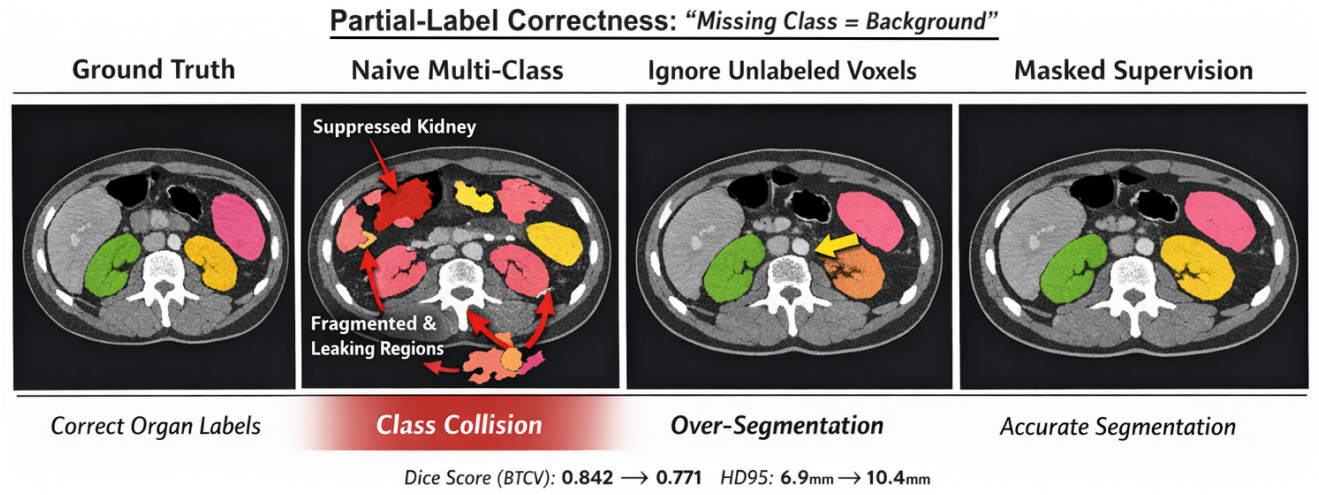


**Figure 8:** Qualitative illustration of partial-label correctness on an abdominal CT slice (BTCV setting). Naïve multi-class supervision (treating missing classes as background) induces class collision, visible as a suppressed kidney and fragmented/leaking predictions. Ignoring unlabeled voxels mitigates false-negative suppression but removes dense negative evidence, leading to over-segmentation (arrow). Masked supervision (ours) preserves correct organ boundaries by applying class-availability masking so that only annotated organs contribute supervised gradients, while missing organs are not treated as negatives.

**Table 2**
Aggregation weight ablation on BTCV in the cross-dataset partial-label federation. We report BTCV 13-organ macro DSC/HD95, plus two organ subsets: (i) overlap organs that appear as labeled targets in some single-organ clients (liver, spleen, pancreas, kidneys), and (ii) BTCV-only organs (e.g., vessels and smaller structures) that are supervised almost exclusively by BTCV. Stability is quantified by the per-round variance of client update norms. Mean±std over three random seeds.

| Aggregation weighting | BTCV (13-org) DSC↑ / HD95↓ | Overlap organs DSC↑ | BTCV-only organs DSC↑ | Major vessels (Ao+IVC) DSC↑ / HD95↓ | $\overline{\mathrm{Var}}_r\ \mathrm{Var}_k(\|\Delta\psi_k^{(r)}\|_2)$ $(\times 10^{-3})\downarrow$ |
|---|---|---|---|---|---|
| Uniform | 0.844±0.019 / 6.8±0.7 | 0.892±0.012 | 0.814±0.026 | 0.853±0.021 / 5.2±0.5 | 3.6±0.6 |
| FedAvg | 0.838±0.021 / 7.4±0.9 | **0.897±0.010** | 0.801±0.031 | 0.838±0.024 / 5.8±0.6 | 4.9±0.8 |
| Label-coverage | 0.848±0.018 / 6.6±0.6 | 0.890±0.013 | 0.823±0.024 | 0.861±0.020 / 5.0±0.4 | 3.0±0.5 |
| **supervision mass (0urs)** | **0.853±0.020 / 6.4±0.7** | 0.889±0.014 | **0.833±0.023** | **0.869±0.019 / 4.8±0.4** | **2.4±0.4** |

dense one-vs-rest supervision for the annotated organs over the full voxel lattice, enabling the model to learn organ-specific decision boundaries while remaining agnostic about organs that lack labels.

### *4.8.2. Impact of Supervision-aware Aggregation*

We isolate the effect of aggregation weighting in the server fusion step (Eqs. (12)- (13)). All variants share the same FedSAM3D backbone (3D adapters + APG), masked supervision (Equation (7)), coupled global-peer distillation, preprocessing/augmentation, and communication schedule. We use the cross-dataset federation with $K$=5 clients, where BTCV is the only client with dense 13-organ supervision, while Spleen/Liver/Pancreas/Kidney clients provide single-organ labels. We report performance only on BTCV (13-organ macro) to directly test whether the dense client is under-represented during aggregation. At round $r$, client weights $\alpha_k^{(r)}$ are set as: (i) **Uniform**: $\alpha_k = 1/|\mathcal{K}^{(r)}|$; (ii) **FedAvg**: $\alpha_k \propto N_k$; (iii) **Label-coverage proxy**: $\alpha_k \propto |\mathcal{O}_k|\, N_k$; (iv) **(Ours) supervision mass**: $\alpha_k \propto s_k$, where $s_k$ estimates the effective supervised signal contributed in a round (Equation (12)), accounting for the number of supervised classes and the supervised voxel lattice under patch-based training. To diagnose optimization stability, we additionally compute the per-round update-norm variance across clients: $\mathrm{Var}_k(\|\Delta\psi_k^{(r)}\|_2)$, averaged over rounds.

Table 2 highlights that aggregation weighting materially affects which organs are retained, when label coverage is heterogeneous. Size-weighted FedAvg yields the best overlap-organ DSC (0.897) but degrades BTCV-only structures (0.801) and increases boundary error (HD95 7.4 mm). This behavior is consistent with a dominance effect: large single-organ datasets contribute frequent and coherent gradients for their labeled organs, and size-weighting amplifies these updates even though they do not provide supervision for most BTCV classes. In contrast, both coverage-aware schemes increase performance on BTCV-only organs, With the proposed supervision-mass weighting providing the greatest improvement on difficult structures (e.g., major vessels: 0.869 DSC / 4.8 mm HD95), suggesting reduced under-representation of dense-label expertise.

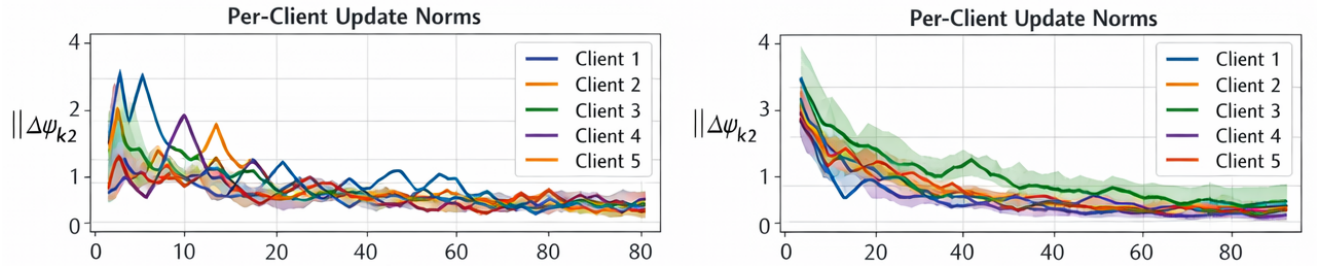


**Figure 9:** Per-client update-norm trajectories across communication rounds under two aggregation schemes. We plot $\|\Delta\psi_k^{(r)}\|_2$ for each client $k$, where $\Delta\psi_k^{(r)} = \psi_k^{(r,E)} - \psi^{(r)}$. Compared with size-weighted FedAvg fusion (left), the proposed supervision-mass weighting (right) yields smoother, more consistently decaying update magnitudes and reduces cross-client dispersion, indicating more stable server trajectories under non-IID data and heterogeneous label coverage. Shaded bands denote variability across runs (mean±std).

**Table 3**
Distillation decomposition on BTCV. We report BTCV macro (13-org) and BTCV pancreas organ performance.

| Variant | BTCV (13-org macro) DSC↑ / HD95↓ (mm) | BTCV Pancreas DSC↑ / HD95↓ (mm) |
|---|---|---|
| Masked sup. only ($\beta$=$\gamma$=0) | 0.836±0.022 / 7.6±0.9 | 0.846±0.050 / 10.0±1.4 |
| + GKD only ($\beta$>0, $\gamma$=0) | 0.846±0.020 / 7.0±0.7 | 0.851±0.046 / **9.1±1.0** |
| + PKD only ($\beta$=0, $\gamma$>0) | 0.842±0.021 / 7.2±0.8 | **0.861±0.042** / 9.6±1.2 |
| Coupled ($\beta$>0, $\gamma$>0) | **0.851±0.019** / **6.6±0.7** | 0.858±0.044 / 9.3±1.1 |

The update-norm variance further indicates that supervised weight updates improve optimization stability. FedAvg exhibits the highest cross-client dispersion ($4.9\times10^{-3}$), implying larger inter-client disagreement in update magnitudes, which is a known precursor to oscillatory server trajectories under non-IID training. Supervision-mass weighting reduces this dispersion by ∼51% relative to FedAvg (2.4 vs 4.9), supporting the claim that weighting by an effective supervised signal better aligns the aggregation step with the true training objective under partial labels. Notably, supervision-mass weighting does not uniformly dominate every subset: Its overlap-organ DSC is slightly lower than FedAvg (0.889 vs 0.897). This suggests a mild shift from overlap-class specialization toward improved coverage of BTCV-only organs, i.e., better organ-wise fairness at the expense of small overlap gains.

Figure 9 visualizes the per-round client update magnitudes $\|\Delta\psi_k^{(r)}\|_2$. With size-weighted aggregation, clients exhibit pronounced early-round spikes and persistent oscillations, reflecting large inter-client disagreement in update scale under non-IID appearance shift and label-space mismatch. In contrast, supervision-mass weighting produces a smoother decay of $\|\Delta\psi_k^{(r)}\|_2$ across all clients and substantially reduces cross-client spread, indicating a more stable server trajectory and better-aligned optimization dynamics when effective supervised signal (rather than dataset size alone) determines fusion weights.

#### *4.8.3. Ablation Study on Distillation Decomposition*

We ablate the proposed knowledge-transfer components to quantify the complementarity of global consistency distillation and peer-guided distillation. All variants use the same FedSAM3D backbone (3D adapters + APG), identical preprocessing/augmentation, the same communication schedule, and masked supervision (Equation (7)). The only change is whether GKD ($\beta$) and/or PKD ($\gamma$) is enabled. We conduct the study on BTCV and report both BTCV 13-organ macro performance and the pancreas organ score to emphasize sensitivity to missing-class transfer on an anatomically challenging, low-contrast target.

Table 3 demonstrates that both transfer pathways contribute distinct benefits, and their coupling yields the most robust overall behavior. With masked supervision only ($\beta$=$\gamma$=0), BTCV macro reaches 0.836 ± 0.022 DSC with 7.6 ± 0.9 mm HD95, while pancreas attains 0.846 ± 0.050 DSC and 10.0 ± 1.4 mm HD95, indicating that partial-label masking alone mitigates missing-as-background collapse but remains susceptible to forgetting and boundary outliers on unlabeled organs under non-IID drift. Enabling GKD ($\beta$>0, $\gamma$=0) improves macro DSC to 0.846 ± 0.020 and reduces macro HD95 to 7.0 ± 0.7 mm, with the strongest effect on pancreas boundary quality (HD95 10.0 → 9.1 mm), consistent with server-teacher guidance acting as a stabilizing semantic prior that suppresses fragmented predictions and reduces rare boundary failures through confidence-gated supervision on missing classes.

In contrast, PKD alone ($\beta$=0, $\gamma$>0) yields a larger gain in pancreas overlap, achieving the best pancreas DSC of 0.861 ± 0.042 (vs. 0.851 ± 0.046 with GKD), while macro DSC improves more modestly to 0.842 ± 0.021. However, PKD does not achieve the lowest pancreas HD95 (9.6 mm vs. 9.1 mm with GKD) and remains slightly worse than the coupled model on macro boundary accuracy (7.2 mm vs. 6.6 mm), suggesting that peer experts more effectively transfer organ-specific competence (improving DSC) but can be less consistent under cross-client appearance shift, where provider-to-target domain mismatch introduces calibration noise that manifests primarily in surface-distance tails.

**Table 4**
Confidence gating ablation on BTCV. Ungated removes the voxel-wise reliability masks for global consistency distillation (GKD; $\mathbf{M}$ in Equation 2) and/or peer-guided distillation (PKD; $\mathbf{M}^{\mathrm{P}}$. For gated settings, thresholds are applied as indicated; when a pathway is ungated, all voxels contribute.

| Variant | $\tau$ (GKD) | $\tau_{\mathrm{P}}$ (PKD) | BTCV macro DSC↑ | BTCV macro HD95↓ | Pancreas DSC↑ | Pancreas HD95↓ |
|---|---|---|---|---|---|---|
| Ungated GKD + ungated PKD | – | – | 0.842±0.022 | 7.4±0.9 | 0.848±0.050 | 10.0±1.3 |
| Gated GKD only (PKD ungated) | 0.7 | – | 0.846±0.021 | 6.9±0.8 | 0.849±0.048 | **9.2±1.2** |
| Gated PKD only (GKD ungated) | – | 0.7 | 0.845±0.022 | 7.2±0.9 | **0.856±0.045** | 9.7±1.2 |
| Both gated (ours; default) | 0.7 | 0.7 | **0.849±0.020** | **6.6±0.7** | 0.854±0.046 | 9.4±1.1 |
| *Threshold sensitivity for both-gated (mean only):* | | | | | | |
| Both gated (sensitivity) | 0.6 | 0.6 | 0.847 | 6.8 | 0.856 | 9.8 |
| Both gated (sensitivity) | 0.8 | 0.8 | 0.846 | 6.7 | 0.849 | **9.1** |

Coupling GKD and PKD ($\beta>0, \gamma>0$) produces the best macro performance, reaching $0.851 \pm 0.019$ DSC and the lowest macro HD95 of $6.6 \pm 0.7$ mm, indicating complementary effects: global distillation regularizes federated semantics and improves stability, while peer transfer injects organ-specialized knowledge that is otherwise diluted by server averaging. Notably, the coupled configuration is not uniformly dominant on every metric—pancreas DSC ($0.858 \pm 0.044$) is marginally below PKD-only ($0.861 \pm 0.042$) which is consistent with a realistic regularization trade-off where enforcing global consistency can slightly temper single-organ overlap gains while still reducing boundary outliers and improving federation-level robustness. These trends collectively support the claim that GKD and PKD address different failure modes (stability/forgetting vs. specialization transfer), and their combination yields the most reliable accuracy–boundary trade-off under partial-label, non-IID federated training.

### *4.8.4. Impact of Confidence Gating for Distillation*

Table 4 evaluates whether distillation benefits arise from reliable knowledge transfer rather than indiscriminate copying of noisy teacher outputs. Removing both reliability masks (ungated GKD + ungated PKD) yields the weakest boundary behavior, with BTCV macro HD95 increasing to $7.4 \pm 0.9$ mm and pancreas HD95 deteriorating to $10.0 \pm 1.3$ mm, consistent with error propagation from low-confidence voxels in anatomically ambiguous regions. Introducing confidence gating on the *global* teacher alone improves boundary robustness more than overlap, reducing BTCV macro HD95 from 7.4 to 6.9 mm and achieving the lowest pancreas HD95 ($9.2 \pm 1.2$ mm), indicating that server guidance is particularly sensitive to cross-domain uncertainty and benefits from conservative filtering. Conversely, gating only the *peer* pathway yields the best pancreas overlap (DSC $0.856 \pm 0.045$) but does not minimize surface error (pancreas HD95 $9.7 \pm 1.2$ mm), suggesting that peer experts can enhance class-specific recall yet still introduce boundary outliers when domain shift is substantial. Enabling both gates provides the best federation-level trade-off (BTCV macro DSC $0.849 \pm 0.020$ and macro HD95 $6.6 \pm 0.7$ mm), while not strictly maximizing pancreas metrics, reflecting a realistic precision-coverage compromise: reliability filtering stabilizes global behavior across organs even if it slightly reduces organ-specific peak gains.

The sensitivity rows further support the reliability–coverage mechanism. A looser threshold ($\tau = \tau_{\mathrm{P}} = 0.6$) marginally increases pancreas DSC (0.856) but worsens pancreas HD95 (9.8 mm), implying that additional low-confidence voxels primarily manifest as surface outliers. A stricter threshold (0.8) improves pancreas boundary accuracy (HD95 9.1 mm) but reduces pancreas DSC (0.849), consistent with discarding moderately confident yet informative voxels. The intermediate threshold (0.7) yields the most balanced macro performance and is therefore used as the default in subsequent experiments.

**Table 5**
Ablation of volumetric adaptation components on BTCV (13-organ macro).

| Variant | DSC ↑ (mean±std) | HD95 ↓ (mm) |
|---|---|---|
| w/o depth positional embedding | 0.846±0.022 | 7.0±0.8 |
| w/o depth mixer | 0.843±0.025 | 7.4±0.9 |
| 3D adapter conv → 2D-only (no depth conv) | 0.852±0.018 | 7.2±0.8 |
| Full volumetric design (ours) | 0.851±0.021 | 6.6±0.7 |

### *4.8.5. Ablation Study on Volumetric Adaptation*

We evaluate the intrinsically volumetric contribution of FedSAM3D by ablating depth-aware components while keeping the federated protocol, masked supervision, Auto Prompt Generator (APG), local optimization budget, and communication schedule fixed. All variants are evaluated on BTCV, reporting macro-averaged performance across 13 organs (DSC/HD95 computed per organ and averaged). We emphasize HD95 because volumetric coherence failures (e.g., inter-slice discontinuities and boundary outliers) often manifest more strongly in surface-distance metrics than in overlap. Table 5 shows that removing depth-aware design choices leads to a consistent increase in HD95, indicating more frequent boundary outliers and reduced through-plane consistency. Notably, the degradation is disproportionately larger in HD95 than in DSC (e.g., full design: 6.6 → 7.0–7.4 mm vs. 0.851 → 0.843–0.846), suggesting that the primary benefit of volumetric adaptation is geometric regularity rather than purely improved foreground occupancy. This behavior is consistent with the common failure mode

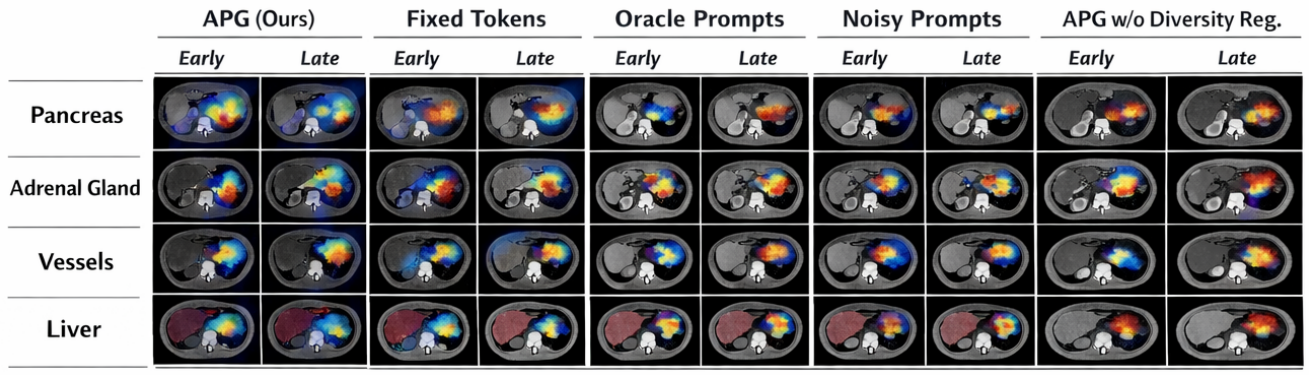


**Figure 10:** Cross-attention analysis of prompting mechanisms in the SAM-style mask decoder. We visualize cross-attention maps (query → volumetric tokens) for four representative organs (pancreas, adrenal gland, vessels, liver) under different prompt variants: APG (ours), fixed learned class tokens, oracle prompts (GT-derived), noisy oracle prompts, and APG without diversity regularization. For each variant, attention from an early and a late decoder block is shown (warm colours indicate closer attention). APG produces progressively sharper and more organ-aligned attention, while fixed tokens remain diffuse; oracle prompts are most localized; noisy prompts degrade localization; and removing diversity regularization reduces organ-specific differentiation.

**Table 6**
Ablation study to evaluate the impact of APG. Metrics are reported as mean±std over three random seeds. BTCV macro is averaged over 13 organs; the pancreas is reported separately.

| Variant | BTCV (13-org macro) | | Pancreas | |
|---|---|---|---|---|
| | DSC ↑ | HD95 ↓ (mm) | DSC ↑ | HD95 ↓ (mm) |
| Fixed learned class tokens (no APG) | 0.842±0.024 | 7.3±0.8 | 0.785±0.056 | 14.1±1.9 |
| Oracle prompts (GT-derived, upper bound) | 0.862±0.018 | 6.0±0.6 | 0.821±0.045 | 11.9±1.5 |
| APG (ours) | **0.851±0.021** | **6.6±0.7** | **0.809±0.048** | **12.5±1.6** |

of 2D-leaning pipelines in volumetric settings: masks remain approximately localized (limited DSC drop) but exhibit slice-to-slice "flicker" and irregular extremities that inflate surface distances.

Among the components, removing the depth mixer causes the largest boundary deterioration (HD95 7.4 ± 0.9 mm), implying that explicit early through-plane feature propagation is critical under patch-based training, where limited context amplifies depth ambiguity and gradient variance. Disabling the depth positional embedding produces a smaller but systematic decline (HD95 7.0 ± 0.8 mm), supporting the interpretation that explicit slice indexing reduces positional aliasing along the depth axis and stabilizes cross-sectional correspondence. Replacing the 3D adapter convolution with a 2D-only operator yields a mixed outcome: the DSC can remain comparable (or even slightly higher within run-to-run variability), yet HD95 deteriorates markedly (7.2 ± 0.8 mm), indicating that depth-aware adapters primarily act as a contour-regularization mechanism that suppresses volumetric boundary artefacts rather than uniformly increasing overlap.

### *4.8.6. Impact of Auto Prompt Generator*

We evaluate the necessity of the proposed Auto Prompt Generator (APG) by isolating whether FedSAM3D's gains stem from genuinely image-conditioned prompting or can be reproduced by trivial per-class tokens. Experiments are conducted on the BTCV multi-organ dataset (13 organs), which is a suitable benchmark for this ablation because it contains both large, high-contrast structures (e.g., liver, spleen) and thin/low-contrast organs (e.g., pancreas, adrenal glands), where prompt quality directly affects boundary placement. All variants use the same volumetric backbone adaptation, masked supervision, federated schedule, and training budget; only the prompt specification mechanism is changed. We report macro-averaged DSC/HD95 across BTCV organs, and additionally highlight the pancreas subset, as thin structures are typically most sensitive to prompt conditioning.

Table 6 indicates that replacing image-conditioned prompting with fixed learned class tokens yields a measurable decline in both overlap and boundary quality, with the effect most pronounced on the pancreas. Specifically, the macro DSC drops from 0.851±0.021 (APG) to 0.842±0.024 (fixed tokens), while HD95 increases from 6.6 ± 0.7 mm to 7.3 ± 0.8 mm, suggesting that non-adaptive prompts introduce organ boundary outliers and reduce volumetric coherence. The pancreas exhibits a larger sensitivity gap: APG improves DSC from 0.785 ± 0.056 to 0.809 ± 0.048 and reduces HD95 from 14.1 ± 1.9 mm to 12.5 ± 1.6 mm. This pattern is consistent with the pancreas being thin, low-contrast, and spatially ambiguous, in which fixed-class embeddings are insufficient to encode scan-specific contextual cues (e.g., patient-dependent appearance and organ adjacency) needed for robust prompt-conditioned decoding.

The oracle prompt setting provides a practical upper bound onhow much prompting can improve segmentation when high-quality prompts are available at both trainingn and test time. Oracle prompts yield the best boundary performance (BTCV HD95 6.0±0.6 mm), but the improvement over APG is modest relative to the gap between APG and fixed tokens, indicating that APG captures a substantial fraction of the benefit of strong prompts without manual

**Table 7**
Effect of stochastic missing-class subsampling $m_k$ for PKD on the Pancreas client. Mean±std over three random seeds. Payload counts the uploaded model update plus the peer-expert transfers received in that round.

| $m_k$ (missing classes/round) | DSC ↑ | HD95 ↓ (mm) | Time/round ↓ (min) | Payload/round ↓ (MB) |
|---|---|---|---|---|
| 1 | 0.803±0.052 | 13.4±2.0 | 14.2±0.4 | 13.0±0.2 |
| 2 | **0.812±0.049** | 12.9±1.8 | 15.6±0.5 | 13.6±0.3 |
| 4 (default) | 0.810±0.046 | 12.6±1.7 | 18.9±0.6 | 14.9±0.3 |
| all missing classes | 0.809±0.050 | **12.4±1.8** | 31.5±0.9 | 19.8±0.4 |

interaction. Importantly, the remaining oracle-APG gap suggests that there is still headroom for improving automatic prompt synthesis, particularly for anatomically variable or weakly contrasted organs. Qualitatively, APG tends to reduce pancreas "fragmentation" and slice-to-slice discontinuities compared to fixed tokens, which aligns with the observed HD95 reduction and supports the claim that APG contributes meaningfully to volumetric prompt reliability rather than acting as a trivial class-identity embedding.

To verify that APG improves performance by providing spatially meaningful, image-conditioned prompts (rather than acting as a class-ID embedding), we visualize mask-decoder cross-attention maps (query → volumetric tokens) for four representative organs. As shown in Figure 10, APG yields attention that progressively sharpens from early to late decoder blocks and concentrates on the ground-truth organ region, particularly for challenging low-contrast structures (pancreas, adrenal gland) and thin targets (vessels). In contrast, fixed learned class tokens induce more diffuse and less organ-specific attention, while oracle prompts produce the tightest localization. Injecting noise into oracle prompts visibly degrades attention localization, and removing the diversity regularizer reduces organ-specific differentiation, indicating prompt collapse. These observations explain the performance gap in Table 6 and directly support FedSAM3D's central claim of auto-prompted, prompt-conditioned decoding.

#### *4.8.7. Impact of Missing-Class Subsampling*

Peer-guided distillation (PKD) can be expensive in partial-label federation because a single-organ client may miss a large fraction of the global label space. To keep peer transfer computationally feasible, FedSAM3D stochastically subsamples a small set of missing classes per round, $\widetilde{\mathcal{O}}_k^{(r)} \subseteq \overline{\mathcal{O}}_k$, with $|\widetilde{\mathcal{O}}_k^{(r)}| = m_k$. We ablate $m_k$ on the pancreas client (single-organ supervision, large missing-class set), where PKD cost is most pronounced and where improvements typically reflect better prevention of forgetting and reduced cross-client bias. All variants use the same backbone, masked supervision, global distillation, confidence gating, and federation schedule; only $m_k$ is changed. We report pancreas DSC/HD95, the average wall-clock time per communication round (end-to-end, including peer inference), and the total communication payload per round (model update + transmitted peer experts; larger $m_k$ increases the number of experts relayed to a client in that round).

Table 7 shows a clear accuracy-efficiency trade-off governed by $m_k$. Increasing $m_k$ from 1 to 2 yields the largest accuracy gain, improving DSC from 0.803 ± 0.052 to 0.812 ± 0.049 while reducing HD95 from 13.4 ± 2.0 mm to 12.9 ± 1.8 mm, with only a modest increase in time/round (14.2→15.6 min) and payload (13.0→13.6 MB). Moving to the default $m_k = 4$ slightly improves boundary quality (HD95 12.6 ± 1.7 mm) but does not consistently increase DSC (0.810±0.046), suggesting diminishing returns once the student receives enough peer signals to regularize missing-class logits. Evaluating an (expensive) "all missing classes" run improves HD95 marginally (to 12.4 ± 1.8 mm) but does not improve DSC (0.809±0.050) and more than doubles the time/round (31.5 min), consistent with over-regularization and residual teacher noise accumulating when a large set of peer constraints is enforced simultaneously. Importantly, the communication cost also increases substantially (to 19.8 MB/round) because more peer experts must be transferred, indicating that exhaustive peer transfer is not cost-effective in cross-silo settings.

### Performance Comparison with SoTA Models

We compare FedSAM3D against five recent state-of-the-art baselines that best align with cross-silo, non-IID volumetric segmentation under fragmented supervision: FednnU-Net [38], UFPS [15], FedIA [43], KD-Synth [19], and FedFMS [29]. All methods share the same federation protocol ($K = 5$), deterministic preprocessing, patch-based training and sliding-window inference, identical communication rounds $R$, and the same evaluation procedure that reports DSC/HD95 only for classes with available ground truth, together with deployment-oriented efficiency indicators. Table 8 reports that FedSAM3D achieves the strongest overall accuracy-efficiency trade-off across all clients. In terms of overlap accuracy, FedSAM3D yields the highest DSC on every dataset: spleen reaches 0.961 ± 0.013 (vs. 0.956±0.015 for KD-Synth and 0.955±0.015 for FedFMS), liver reaches 0.973 ± 0.010 (vs. 0.970 ± 0.012 for KD-Synth), pancreas reaches 0.862±0.041 (vs. 0.846±0.044 for FedFMS), kidney reaches 0.956±0.015 (vs. 0.949±0.018 for FedFMS), and the BTCV 13-organ macro reaches 0.851 ± 0.021 (vs. 0.838 ± 0.023 for FedFMS). Notably, the largest absolute DSC margin is observed on the pancreas client (+0.016 over FedFMS and +0.020 over KD-Synth), consistent with the pancreas being both anatomically challenging (thin, low-contrast boundaries) and particularly vulnerable to cross-client supervision-mismatch effects in partial-label federation.

Boundary quality exhibits a similarly consistent pattern. FedSAM3D obtains the lowest HD95 on all benchmarks, reducing surface error to 2.9 mm for spleen, 3.5 mm for liver, 8.8 mm for pancreas, 3.8 mm for kidney, and 6.6 mm for BTCV. Relative to the best competing baseline per dataset, FedSAM3D provides non-trivial boundary improvements, e.g., on BTCV HD95 improves from 7.2 mm (FedFMS) to 6.6 mm, and on pancreas from 9.7 mm (FedFMS) to 8.8 mm. These gains indicate that the proposed coupled

**Table 8**
SoTA comparison across the cross-dataset federation. Dice (DSC) and HD95 are reported only for labeled classes at each client and macro-averaged for BTCV. Payload and peak GPU memory quantify deployment efficiency under parameter-efficient federated training.

| Method | DSC ↑ (mean±std) / HD95 ↓ (mm) | | | | | Payload (MB/round) ↓ / Peak Mem (GB) ↓ |
|---|---|---|---|---|---|---|
| | Spleen | Liver | Pancreas | Kidney | BTCV (13-org macro) | |
| FednnU-Net [38] | 0.946±0.018 / 3.9 | 0.962±0.014 / 4.5 | 0.812±0.052 / 11.8 | 0.934±0.021 / 5.2 | 0.802±0.028 / 8.7 | 86.0 / 19.6 |
| UFPS [15] | 0.951±0.017 / 3.6 | 0.966±0.013 / 4.2 | 0.827±0.048 / 10.9 | 0.941±0.020 / 4.8 | 0.815±0.026 / 8.1 | 72.4 / 20.1 |
| FedIA [43] | 0.953±0.016 / 3.4 | 0.968±0.012 / 4.0 | 0.834±0.047 / 10.5 | 0.944±0.019 / 4.6 | 0.821±0.025 / 7.9 | 73.1 / 20.0 |
| KD-Synth [19] | 0.956±0.015 / 3.2 | 0.970±0.012 / 3.8 | 0.842±0.045 / 9.9 | 0.948±0.018 / 4.3 | 0.832±0.024 / 7.4 | 28.6 / 20.8 |
| FedFMS [29] | 0.955±0.015 / 3.3 | 0.969±0.012 / 3.9 | 0.846±0.044 / 9.7 | 0.949±0.018 / 4.2 | 0.838±0.023 / 7.2 | 14.9 / 21.7 |
| **FedSAM3D (ours)** | **0.961±0.013 / 2.9** | **0.973±0.010 / 3.5** | **0.862±0.041 / 8.8** | **0.956±0.015 / 3.8** | **0.851±0.021 / 6.6** | **12.4 / 21.3** |

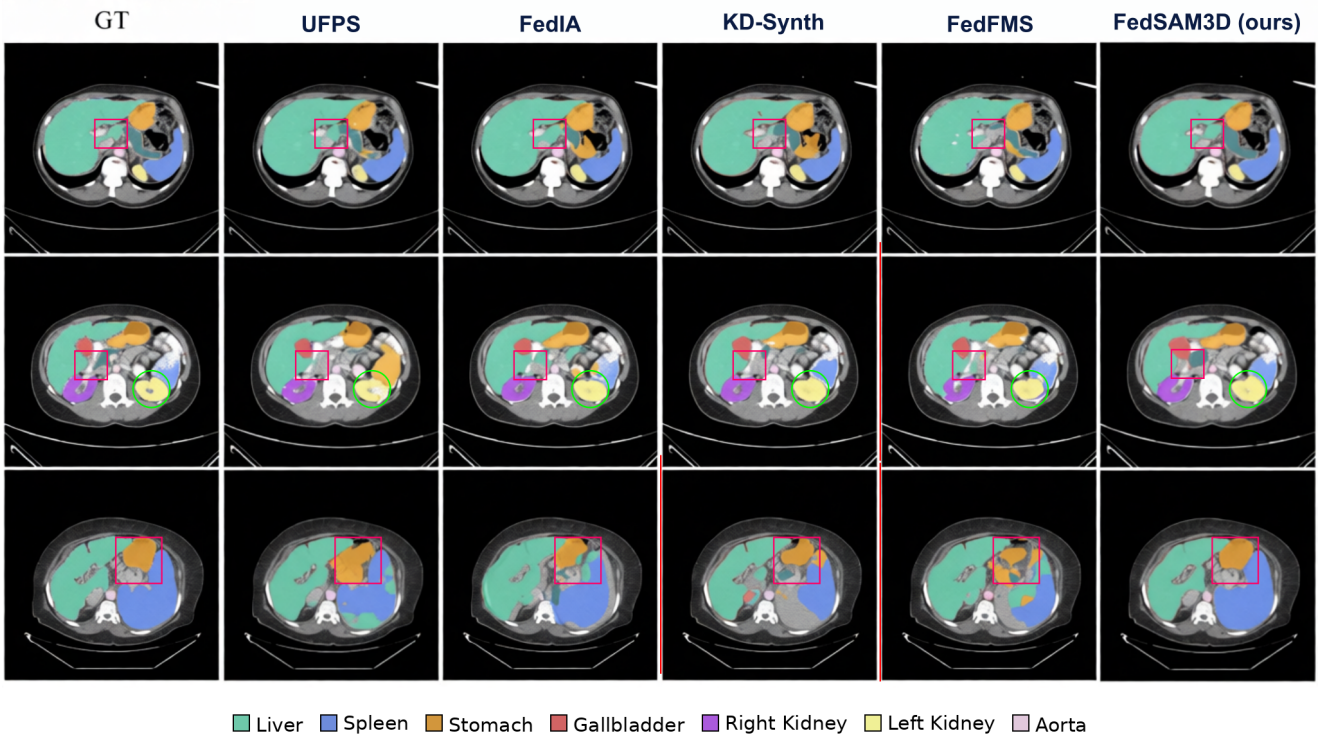


**Figure 11:** Qualitative comparison on BTCV dataset Representative axial slices from the BTCV dataset The overlays visualize major abdominal organs. Magenta boxes highlight challenging inter-organ interfaces where small boundary deviations cause large surface-distance penalties, and the green circle marks a localized failure case around the left-kidney region. FedSAM3D better preserves boundary placement and reduces leakage/class confusion, producing contours that are most consistent with the ground truth.

transfer (server-level consistency plus peer specialization) is not merely increasing volumetric overlap but also improving contour fidelity, which is critical for downstream clinical tasks sensitive to boundary placement (e.g., dosimetry planning and organ-at-risk delineation).

x Qualitative evidence on BTCV. The numerical HD95 improvements are reflected visually in Figure 11, which compares representative BTCV axial slices across four strong baselines and FedSAM3D. While UFPS/FedIA and KD-Synth often recover the coarse organ extent, they exhibit noticeable boundary bias in challenging interface regions (magenta boxes), including mild leakage and contour smoothing that shifts organ boundaries. FedFMS further shows fragmented or mixed predictions in the most ambiguous slice. In contrast, FedSAM3D yields cleaner inter-organ separation and more continuous, anatomically consistent contours, particularly in the highlighted regions and the left-kidney failure case (green circle), aligning with its lowest HD95 on BTCV in Table 8.

Results highlight how FedSAM3D behaves under distinct axes of heterogeneity. First, under label-space fragmentation (single-organ clients), UFPS and FedIA improve upon FednnU-Net (e.g., pancreas DSC: 0.827 → 0.834), confirming the importance of partial-label-aware learning. However, FedSAM3D further advances performance (pancreas DSC: 0.862; HD95: 8.8 mm) by combining masked supervision with two complementary missing-class transfer routes:

global distillation regularizes unlabeled organs toward a federation-wide semantic prior, while peer-guided distillation injects organ-specific expertise that may be diluted by global averaging. Second, under non-IID appearance shift (cross-dataset federation), the improvements over KD-Synth suggest that direct peer specialization transfer can be advantageous when proxy-domain mismatch or dataset-specific biases limit teacher reliability. Third, relative to FedFMS, the gains indicate that parameter-efficient foundation-model tuning alone is insufficient in this setting unless coupled with intrinsically volumetric adaptation and automatic class-conditioned prompting; FedSAM3D's APG-driven prompting and 3D-aware adapters better preserve volumetric coherence.

From an efficiency perspective, FedSAM3D is communication efficient while sustaining competitive GPU memory requirements. The per-round payload is reduced to 12.4 MB, which is $\sim 7\times$ smaller than KD-Synth (28.6 MB) and nearly an order of magnitude smaller than FednnU-Net (86.0 MB), while remaining comparable to FedFMS (14.9 MB). Peak GPU memory is 21.3 GB, comparable to other SAM-style/transformer-based baselines (20.8-21.7 GB) and slightly higher than FednnU-Net (19.6 GB), reflecting the additional volumetric attention and APG computation.

### 4.9. Discussion

FedSAM3D targets a practically dominant but underaddressed regime for volumetric segmentation: cross-silo federation with simultaneous non-IID appearance shift and fragmented organ supervision. Across all clients, FedSAM3D yields the best DSC and lowest HD95 (Table 8), and the qualitative BTCV comparisons (Figure 11) show that these gains are not limited to bulk foreground occupancy. Competing methods frequently recover the coarse organ extent yet exhibit interface-specific boundary bias and localized leakage in challenging multi-organ junctions (magenta boxes) and around the kidney region (green circle), which manifests as surface-distance outliers and degraded HD95.

Ablations on supervision strategies (Table 1 and Figure 8) confirm that naive multi-class learning in partial-label federation introduces systematic false-negative gradients by treating missing organs as background that produce class collision and organ suppression after aggregation. While ignoring unlabeled voxels mitigates this collapse, it also discards dense negative evidence and weakens one-vs-rest constraints, leading to over-segmentation and boundary leakage. In contrast, class-availability masking preserves dense supervision for annotated organs across the full voxel lattice while remaining agnostic to unlabeled targets, producing the most stable and anatomically consistent predictions. It shows that partial-label handling is not merely a data-efficiency issue but a correctness constraint on the federated objective.

When one client (BTCV) provides dense supervision while others are single-organ cohorts, server fusion must reflect effective labeled signal rather than dataset size alone. The aggregation ablation (Table 2 and Figure 9) shows that size-weighted FedAvg overemphasizes large single-organ clients, improving overlap on organs that receive frequent supervision but degrading BTCV-only structures (e.g., vessels) and increasing HD95, consistent with underrepresentation of dense multi-class expertise. Weighting by supervision mass $s_k$ improves BTCV macro performance and reduces update-norm variance, indicating smoother server trajectories under non-IID shift. This suggests that label coverage heterogeneity changes the meaning of "client contribution": in multi-organ federation, the relevant quantity is not the number of volumes per site but the amount of supervised organ-voxel evidence that aligns with the global objective.

Distillation decomposition (Table 3) supports a division of labor between the two transfer routes. GKD improves macro boundary robustness and notably reduces pancreas HD95, consistent with a semantic regularizer that suppresses fragmented predictions and mitigates forgetting on unlabeled organs. PKD provides stronger organ-specific overlap improvements (notably pancreas DSC), consistent with transferring specialized competence that can be diluted by global averaging. The coupled configuration achieves the best macro DSC/HD95 trade-off, indicating that stability-oriented global guidance and specialization-oriented peer transfer address different failure modes and interact constructively. Importantly, confidence gating (Table 4) is critical to prevent propagation of unreliable teacher signals under domain shift that distillation in non-IID volumetric settings must be reliability-aware to avoid amplifying localized errors.

FedSAM3D's gains cannot be attributed solely to adopting a SAM-style backbone. Ablations on depth-aware components (Table 5) show that removing depth positional encoding or depth mixing consistently increases HD95, indicating more boundary outliers and reduced volumetric coherence even when DSC changes are modest. Likewise, the APG study (Table 6 and Figure 10) indicates that image-conditioned prompting is necessary for robust multi-organ performance: fixed learned class tokens degrade both DSC and HD95, particularly on the pancreas, while oracle prompts provide an upper bound that APG approaches without manual interaction.

FedSAM3D reduced communication payload (Table 8), demonstrating that parameter-efficient tuning is compatible with strong cross-silo performance when combined with supervision-correct objectives and knowledge transfer. The missing-class subsampling study (Table 7) further clarifies that peer transfer can be made computationally feasible without exhaustive distillation across all missing organs each round; moderate subsampling attains most of the accuracy benefit with limited increases in time and payload, while exhaustive transfer exhibits diminishing returns.

### 4.10. Limitations

Despite high performance of FedSAM3D for cross-institutional multi-organ segmentation under partial-label supervision, several limitations remain. The framework

assumes a fixed global organ vocabulary $O$ (the union of all annotated structures across sites), which precludes zero-shot generalization to unseen anatomical structures and necessitates federation-wide retraining or vocabulary expansion when new organs/pathologies appear. Although parameter-efficient adaptation reduces communication relative to full-model exchange, peer-guided distillation scales poorly with federation size $K$ (effectively quadratic peer dissemination). Missing-class subsampling ($m_k \ll |\tilde{O}_k|$) lowers per-round cost but remains prohibitive for large federations (e.g., $K > 20$) and/or large vocabularies (e.g., $C > 20$). The method is evaluated only in single-modality CT federations; cross-modality transfer (e.g., MRI→CT) is not addressed, and the adapters lack explicit mechanisms to reconcile fundamentally different contrast formation processes. The approach further assumes expert-quality labels and does not model inter-rater variability or site-specific annotation biases, allowing systematic label noise to propagate through distillation despite confidence gating. Empirical validation is limited to relatively static abdominal anatomy, leaving performance on highly deformable regions and dynamic 4D imaging untested.

### 4.11. Future Work Directions

Key directions include: (i) cross-modality federation via modality-invariant representation learning (e.g., contrastive alignment across CT/MRI/US without paired acquisitions); (ii) integration of differential privacy through noise-calibrated, sparse update mechanisms to mitigate inversion risks while preserving partial-label performance; (iii) temporal extensions for 4D imaging using recurrent or transformer-based temporal encoders to enforce cross-time consistency; (iv) uncertainty-aware distillation that propagates epistemic/aleatoric uncertainty (e.g., MC dropout or evidential objectives) to improve clinical reliability; (v) self-supervised pre-adaptation of the SAM backbone on unlabeled multi-site data to improve robustness under extreme label sparsity; (vi) longitudinal adaptation modules for evolving anatomy (e.g., tumor progression, post-surgical change) while maintaining global compatibility; and (vii) hardware-aware deployment (e.g., quantization-aware training co-designed with adapters) to enable efficient edge inference in resource-constrained settings.

## 5. Conclusion

This work present a federated learning framework designed to address the critical challenges of volumetric multi-organ segmentation under realistic cross-silo constraints: non-IID imaging distributions and fragmented, partial-label supervision. By integrating a parameter-efficient volumetric adaptation of a SAM-style foundation model with a partial-label-aware federated objective and a novel coupled distillation strategy, the framework enables the collaborative learning of a single, robust 3D promptable segmenter without centralizing data or requiring full annotation overlap across institutions. The core design contributions are fourfold. First, the parameter-efficient 3D adaptation, featuring depth-aware modules and APG, extends the benefits of prompt-conditioned decoding to volumetric data while maintaining communication and memory efficiency. Second, the explicit masked-supervision objective prevents the systematic error of treating locally unannotated organs as background, a fundamental issue in partial-label federation. Third, the supervision-aware aggregation scheme weights client updates by their effective supervised signal, stabilizing optimization when label coverage is heterogeneous. Fourth, the coupled global-peer distillation mechanism provides a dual pathway for knowledge transfer: global consistency distillation enforces federation-wide semantic alignment to mitigate forgetting, while peer-guided distillation injects organ-specialized expertise to counteract the dilution effects of averaging.

Evaluation on a cross-dataset federation of abdominal CT benchmarks demonstrated that FedSAM3D high performance on volumetric overlap and boundary accuracy for low-contrast structures. Relative to recent federated segmentation and foundation-model SoTA models, FedSAM3D achieve high accuracy-boundary trade-off on every client, attaining DSC/HD95 of 0.961/2.9 (spleen), 0.973/3.5 (liver), 0.862/8.8 (pancreas), 0.956/3.8 (kidney), and a BTCV 13-organ macro of 0.851/6.6, while maintaining a low communication payload of 12.4 MB/round. Ablation studies show that masked supervision is crucial to avoid class collisions; the coupled distillation strategy provided superior stability and transfer of specialization compared to either pathway alone; and the volumetric adaptation components were essential for achieving anatomically coherent 3D segmentations. Furthermore, the framework maintained practical efficiency and reduced communication payloads through parameter-efficient tuning.

## References


[1] Tareq Mahmod AlZubi and Hamza Mukhtar. Federated knowledge distillation with 3d transformer adaptation for weakly labeled multi-organ medical image segmentation. *IEEE Access*, 2025.

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[3] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[4] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.

[5] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[6] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] M-P Dubuisson and Anil K Jain. A modified hausdorff distance for object matching. In *Proceedings of 12th international conference on pattern recognition*, volume 1, pages 566–568. IEEE, 1994.

[9] Jia Fu, He Li, Tao Lu, Shaoting Zhang, and Guotai Wang. Um-sam: Unsupervised medical image segmentation using knowledge distillation from segment anything model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 616–626. Springer, 2025.

[10] Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey. *Pattern recognition*, 151:110424, 2024.

[11] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[12] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[13] Lina Huang, Yixiong Liang, and Jianfeng Liu. Des-sam: Distillation-enhanced semantic sam for cervical nuclear segmentation with box annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 223–234. Springer, 2024.

[14] Ziyan Huang, Jin Ye, Haoyu Wang, Zhongying Deng, Zhikai Yang, Yanzhou Su, Jie Liu, Tianbin Li, Yun Gu, Shaoting Zhang, et al. Revisiting model scaling with a u-net benchmark for 3d medical image segmentation. *Scientific Reports*, 15(1):29795, 2025.

[15] Le Jiang, Li Yan Ma, Tie Yong Zeng, and Shi Hui Ying. Ufps: A unified framework for partially annotated federated segmentation in heterogeneous data distribution. *Patterns*, 5(2), 2024.

[16] Xixi Jiang, Dong Zhang, Xiang Li, Kangyi Liu, Kwang-Ting Cheng, and Xin Yang. Labeled-to-unlabeled distribution alignment for partially-supervised multi-organ medical image segmentation. *Medical Image Analysis*, 99:103333, 2025.

[17] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[18] Julien Khlaut, Elodie Ferreres, Daniel Tordjman, Helene Philippe, Tom Boeken, Pierre Manceron, and Corentin Dancette. Radsam: Segmenting 3d radiological images with a 2d promptable model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 435–444. Springer, 2025.

[19] Soopil Kim, Heejung Park, Philip Chikontwe, Myeongkyun Kang, Kyong Hwan Jin, Ehsan Adeli, Kilian M Pohl, and Sang Hyun Park. Communication efficient federated learning for multi-organ segmentation via knowledge distillation with image synthesis. *IEEE Transactions on Medical Imaging*, 2025.

[20] Soopil Kim, Heejung Park, Philip Chikontwe, Myeongkyun Kang, Kyong Hwan Jin, Ehsan Adeli, Kilian M Pohl, and Sang Hyun Park. Communication efficient federated learning for multi-organ segmentation via knowledge distillation with image synthesis. *IEEE Transactions on Medical Imaging*, 2025.

[21] Soopil Kim, Heejung Park, Myeongkyun Kang, Kyong Hwan Jin, Ehsan Adeli, Kilian M Pohl, and Sang Hyun Park. Federated learning with knowledge distillation for multi-organ segmentation with partially labeled datasets. *Medical image analysis*, 95:103156, 2024.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[23] Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, page 12. Munich, Germany, 2015.

[24] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[25] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

[26] Junjie Liang, Peng Cao, Wenju Yang, Jinzhu Yang, and Osmar R Zaiane. 3d-sautomed: Automatic segment anything model for 3d medical image segmentation from local-global perspective. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer, 2024.

[27] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363, 2020.

[28] Liyang Liu, Zihan Wang, Minh Hieu Phan, Bowen Zhang, Jinchao Ge, and Yifan Liu. Bpkd: Boundary privileged knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1062–1072, 2024.

[29] Yuxi Liu, Guibo Luo, and Yuesheng Zhu. Fedfms: Exploring federated foundation models for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 283–293. Springer, 2024.

[30] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[31] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82:102616, 2022.

[32] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[33] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[34] Alessio Mora, Irene Tenison, Paolo Bellavista, and Irina Rish. Knowledge distillation for federated learning: a practical guide. *arXiv preprint arXiv:2211.04742*, 2022.

[35] Hamza Mukhtar, Adil Afzal, Sultan Alahmari, and Saud Yonbawi. Ccgn: Centralized collaborative graphical transformer multi-agent reinforcement learning for multi-intersection signal free-corridor. *Neural networks*, 166:396–409, 2023.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] Grzegorz Skorupko, Fotios Avgoustidis, Carlos Martín-Isla, Lidia Garrucho, Dimitri A Kessler, Esmeralda Ruiz Pujadas, Oliver Díaz, Maciej Bobowicz, Katarzyna Gwoździewicz, Xavier Bargalló, et al. Federated nnu-net for privacy-preserving medical image segmentation. *Scientific Reports*, 15(1):38312, 2025.

[39] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC*

*medical imaging*, 15(1):29, 2015.

[40] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022.

[41] Zhikai Wei, Wenhui Dong, Peilin Zhou, Yuliang Gu, Zhou Zhao, and Yongchao Xu. Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–543. Springer, 2024.

[42] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.

[43] Yangyang Xiang, Nannan Wu, Li Yu, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. Fedia: Federated medical image segmentation with heterogeneous annotation completeness. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 373–382. Springer, 2024.

[44] Xuanang Xu, Hannah H Deng, Jamie Gateno, and Pingkun Yan. Federated multi-organ segmentation with inconsistent labels. *IEEE transactions on medical imaging*, 42(10):2948–2960, 2023.

[45] Zhou Zheng, Yuichiro Hayashi, Masahiro Oda, Takayuki Kitasaka, Kazunari Misawa, and Kensaku Mori. Federated 3d multi-organ segmentation with partially labeled and unlabeled data. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–14, 2024.

[46] Jiaxin Zhuang, Linshan Wu, Xuefeng Ni, Xi Wang, Liansheng Wang, and Hao Chen. Bio2vol: Adapting 2d biomedical foundation models for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer, 2025.