# TRIGGERMATCH: CONDITIONAL SEMANTIC ACQUISITION AND MAINTENANCE FOR SEMI-SUPERVISED 3D MEDICAL IMAGE SEGMENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Conventional semi-supervised multi-class semantic segmentation (SSS) treats the role of unlabeled data as identical to that of labeled data: both are expected to increase the model's semantic understanding. We recast SSS as *semantic acquisition* and *semantic maintenance*, two responsibilities that are assumed precisely by learning from labeled data and learning from unlabeled data, respectively. Based on this perspective, we introduce TriggerMatch, which partially masks the training process of semantic acquisition, so that acquisition no longer duplicates what maintenance already preserves, while the maintenance process itself receives richer gradient diversity. Meanwhile, learning from labeled data is supervised by a correctness metric computed against ground truth, and this metric governs the acquisition and maintenance states of the corresponding semantic classes. Furthermore, we conduct extensive ablation and parameter studies to verify the effectiveness of our design, and benchmark TriggerMatch on three standard 3D medical image datasets: AMOS2022, FLARE2022, and Synapse, where it achieves state-of-the-art performance.

## 1 INTRODUCTION

Deep learning now drive volumetric medical image analysis, yet their supervised formulation collides with the economics of clinical data: delineating abdominal organs in CT demands radiologist-level expertise, and annotating a single three-dimensional volume takes tens of minutes to hours (Ji et al., 2022; Wasserthal et al., 2023). Semi-supervised segmentation (SSS) reduces annotation demands by learning from a small labeled set together with a much larger pool of unlabeled data, making it particularly appealing for volumetric medical image segmentation, where dense annotations are expensive to obtain. Beyond annotation cost, medical images differ structurally from the natural RGB images on which most SSL methods were developed. CT abdominal data are inherently single-channel intensity volumes, and the depicted anatomy follows a shared body plan: across patients, scanners, and even different datasets, homologous organs occupy a geometrically consistent arrangement, exhibit similar anatomical appearance, and appear in similar spatial contexts. This cross-dataset geometric, anatomical, and contextual regularity supplies exactly the structure that the smoothness and cluster assumptions underlying SSL rely on, so an SSL framework that exploits it effectively converts archived unlabeled clinical volumes into a training resource and substantially lowers the barrier for deploying medical AI at scale.

Pseudo-labeling and consistency regularization are the two dominant mechanisms of modern SSL. Pseudo-labeling emulates supervised learning on unlabeled data: the model generates masks from its own predictions to stand in for ground-truth annotations, thereby constructing a self-training signal (Scudder, 1965; McLachlan, 1975). Consistency regularization instead requires the model to produce identical predictions for the same input under different augmentations, with the aim of improving its understanding of the underlying data semantics (Tarvainen & Valpola, 2017; Chen et al., 2021; 2023). Both mechanisms, however, are consumers of semantics rather than producers. Pseudo-labeling presupposes that the model already possesses a basic semantic understanding: in its absence, training on self-generated labels merely reinforces the model's erroneous decision boundaries, corrupting the semantic space it has already established rather than driving any further fit to the data. The degenerate case is unambiguous: a model that has never been fitted to the train-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ing data cannot possibly acquire any semantic understanding from labels generated by itself or by any similarly uninformed model. Consistency regularization rests on the same presupposition; it improves generalization only when the model can already recognize context and class semantics, and what it refines is the detail of that understanding rather than the capability itself. Theoretical analyses of self-training formalize this circularity: without a semantic prior established by labeled supervision or transferred through pretraining, the two mechanisms cannot ground the very classes they are asked to propagate (Oymak & Gülcü, 2021; Wei et al., 2021). Their benefit therefore rests on semantic capability accumulated beforehand. This dependence is compounded by the modality itself. CT provides a single intensity channel, so the strong views of a weak-strong pipeline can be built only from geometric and intensity transforms; these perturbations yield far less view diversity than the color-space augmentations of natural-image pipelines, and the consistency gap that powers FixMatch-like training consequently narrows considerably on abdominal CT.

Meanwhile, abdominal multi-organ data enjoys a degree of geometric similarity and anatomical consistency that natural images rarely afford (He et al., 2023; Wu et al., 2024; Iglesias & Sabuncu, 2015; Wasserthal et al., 2023), making it inherently amenable to SSS: within a class, unlabeled data resemble their labeled counterparts far more closely than natural images resemble one another. Under the prevailing convention of randomly initialized backbones, the model enters training with essentially no prior semantics, so the scarce labeled data are the only material from which class semantics can first be formed. We therefore decompose medical SSS into two asymmetric responsibilities: semantics unlearned by the model are acquired exclusively from the labeled data, whereas the unlabeled data serves to maintain the semantics already formed. Guided by this decoupling, we propose TriggerMatch, which lets the model acquire the semantics of every class faster and more uniformly across classes.

In summary, our key contributions are:

- We reformulate multi-class SSS as a dynamic allocation problem between semantic acquisition and semantic maintenance, motivated by heterogeneous class-wise learning dynamics.
- We propose TriggerMatch, a reversible class-wise supervision routing framework that dynamically withdraws and restores labeled supervision according to class-wise learning status.
- We provide theoretical and empirical validation of dynamic supervision allocation. Extensive experiments on AMOS2022, FLARE2022, and Synapse demonstrate consistent improvements over strong semi-supervised segmentation baselines.

## 2 Related Work

**Semi-Supervised Medical Image Segmentation.** Consistency regularization enforces invariant predictions under perturbed views and temporal ensembling (Laine & Aila, 2017; Tarvainen & Valpola, 2017). Its medical applications gate the consistency target by uncertainty (Yu et al., 2019), enforce shape or dual-task priors (Li et al., 2020; Luo et al., 2021a), and rectify multi-scale predictions (Luo et al., 2021b). Cross pseudo supervision (Chen et al., 2021) treats two models as equals: the output of each serves as the pseudo-label that supervises the other, while the predictions of the two models are expected to remain as consistent as possible. Gradient-aware reweighting rescales per-class gradients to counteract class imbalance (Qi et al., 2024).

**Confidence-Gated Pseudo-Labeling.** FixMatch (Sohn et al., 2020) admits unlabeled supervision only where the model exceeds a fixed confidence threshold. FreeMatch (Wang et al., 2023) replaces the constant with a self-adaptive class-wise threshold, and SoftMatch (Chen et al., 2023) trades pseudo-label quantity against quality through a soft weighting function. ReFixMatch (Nguyen & Yang, 2023) rectifies low-confidence predictions against class prototypes before admitting them, while SAFixMatch (Li et al., 2025) attributes FixMatch's generalization advantage to the smoothness of its threshold schedule and smooths the schedule further. All of these methods modulate only the unlabeled loss, and confirmation bias—the amplification of early errors that self-gating invites—is their characteristic failure mode for the very classes the gate starves. The labeled objective remains at full strength for every class irrespective of how the confidence gate treats that class, so a class whose predictions rarely clear the gate receives progressively less unlabeled signal while its labeled gradients saturate; the two objectives interact only implicitly through shared weights. Ex-

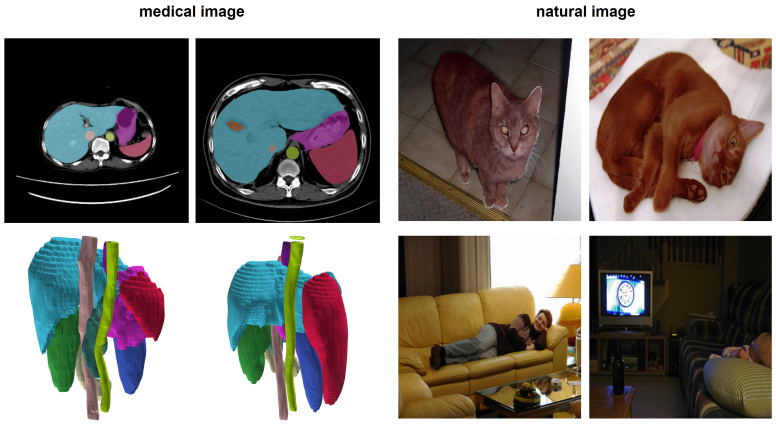



Figure 1: Intra-class appearance variance in the two domains. **Left:** segmented abdominal anatomy from two different patients: homologous organs agree in shape, spatial arrangement, and intensity, so a class forms a tight appearance cluster. **Right:** two instances of a natural-image class (cats) under different poses, backgrounds, and illumination.

isting methods leave this asymmetry unaddressed: mature classes are never retired from the labeled objective, while classes starved by the gate receive no compensating signal.

## 3 METHOD

This section is organized around one question: which supervision should each semantic class receive at each stage of training? Section 3.1 formalizes the task and the static objective shared by current semi-supervised segmenters; Section 3.2 re-examines that objective through the lens of class-wise learning dynamics; Section 3.3 delineates what unlabeled data can and cannot contribute; and Section 3.4 presents the TriggerMatch component. Calligraphic letters denote sets and losses ($\mathcal{D}$, $\mathcal{L}$, $\mathcal{V}$), subscripts $l$ and $u$ mark the labeled and unlabeled branches, $c$ indexes foreground classes, and $t$ indexes optimization steps.

### 3.1 PROBLEM SETUP

We consider SSS with a labeled set $\mathcal{D}_l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ and an unlabeled set $\mathcal{D}_u = \{x_j^u\}_{j=1}^{N_u}$. For an unlabeled volume we construct a weakly augmented view $a_w(x^u)$ and a strongly augmented view $a_s(x^u)$, with the specific augmentation recipes given in Appendix A. Both the FixMatch-like static baseline and TriggerMatch train a student network $f_\theta$ while maintaining an exponential-moving-average (EMA) teacher $f_{\bar\theta}$ of identical architecture, a common configuration in SSS (Tarvainen & Valpola, 2017). After student step $t$, the teacher is updated as

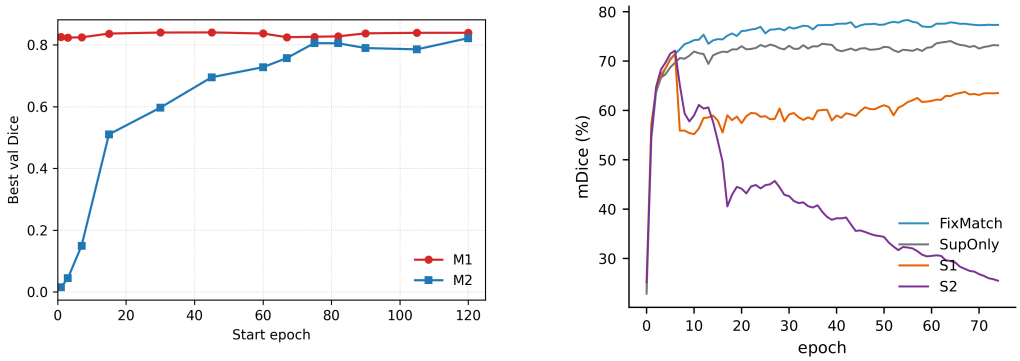
$$\bar\theta_t = \alpha_t\,\bar\theta_{t-1} + (1-\alpha_t)\,\theta_t, \qquad \alpha_t = \min\!\big(1-\tfrac{1}{t},\ \alpha_{\max}\big)\,, \tag{1}$$

where $\alpha_{\max} = 0.996$. The teacher supplies temporally smoothed predictions: they serve as pseudo-label targets on unlabeled data below, and as the class-wise evaluator on labeled data in Section 3.4. The objective shared by current semi-supervised segmenters is *static*: every class is trained by both the labeled and the unlabeled loss at every step.

$$\mathcal{L}_{\text{static}} = \mathcal{L}_l + \lambda_u\,\mathcal{L}_u. \tag{2}$$

The labeled term $\mathcal{L}_l$ is the standard supervised segmentation loss with equal weights: a voxel-wise cross-entropy averaged over all labeled voxels, plus the per-class Dice averaged over the foreground classes. The unlabeled term follows the FixMatch (Sohn et al., 2020): the teacher predicts the weak view and generates pseudo-labels, and the student is supervised on the strong view,

$$\mathcal{L}_u = \frac{1}{|\mathcal{V}_u|}\sum_{v\in\mathcal{V}_u} \mathbb{I}[s_v^u \ge \tau_u]\ \mathrm{CE}(f_\theta(a_s(x^u))_v,\ \tilde y_v)\,, \tag{3}$$




(a) (b)

Figure 2: **(a)** Best validation Dice when one loss term of the static objective is disabled from the indicated epoch onward: **(M1)** disables the unlabeled loss, **(M2)** the labeled loss. **(b)** Validation mDice when class-wise supervision is rerouted after warm-up: **(SupOnly)** uses labeled data only; **(FixMatch)** is the static baseline; **(S1)** directs labeled supervision to the acquisition set $\mathcal{C}_{\text{acq}}^{(t)}$ and unlabeled supervision to the maintenance set $\mathcal{C}_{\text{maint}}^{(t)}$ only; **(S2)** is the inverse routing. Unlabeled supervision alone cannot produce new semantic understanding (quantified in Section 4.3).

where $\tilde{y}_v = \arg\max_c \text{softmax}(f_{\bar{\theta}}(a_w(x^u)))_v$, $s_v^u$ is the confidence of that prediction, and $\mathcal{V}_u$ collects the voxels of the unlabeled data. Two conventions of Eq. (3) matter for what follows. First, the threshold $\tau_u$ gates voxels, not classes: the unlabeled signal reaching a class depends on how confidently the model already predicts it. Second, the denominator counts all voxels rather than only the selected ones, so the magnitude of $\mathcal{L}_u$ scales with pseudo-label coverage. Both properties shape the analysis of Section 3.3.

TriggerMatch departs from Eq. (2) along a single degree of freedom. It maintains a binary, class-wise, time-varying state $z_{c,t} \in \{0, 1\}$, constructed in Section 3.4, that decides whether the *positive anchors* of class $c$ participate in $\mathcal{L}_l$. Setting $z_{c,t} \equiv 0$ for every foreground class and every step, together with $z_{0,t} \equiv 0$ for the background by construction, recovers Eq. (2) exactly. TriggerMatch leaves the unlabeled branch of Eq. (3). Measured differences against the static baseline are therefore attributable to routing labeled supervision alone, and the protocol of Section 4 holds all shared components fixed across the methods we compare.

### 3.2 Rethinking Semi-Supervised Medical Image Segmentation

In a static baseline, the per-class learning curves fan out early and remain separated. The static objective ignores this heterogeneity: it grants every class identical labeled anchoring and identical eligibility for unlabeled supervision at every step. Four facts, all measured on static baselines, indicate that this uniformity misallocates supervision on both sides.

**(F1) Classes complete acquisition at different times.** The gap between the earliest- and latest-acquired classes spans most of the training schedule.

**(F2) Confidence reads out what the model already understands.** As training proceeds, per-class confidence trends upward, and a voxel's confidence tends to track its correctness. Both trends follow from the training objective itself, since gradient descent on the cross-entropy loss drives the predicted probability of the correct class toward one on every voxel it supervises.

**(F3) Unlabeled training maintains, but does not acquire, class semantics.** As Figure 2 shows, once labeled supervision is removed, training on unlabeled data alone scarcely improves semantic understanding.

**(F4) Medical images exploit unlabeled data more readily than natural images.** Unlike camera imagery, CT intensity is a calibrated physical quantity: each voxel value encodes the attenuation of the underlying tissue, so air-level values denote background and abdominal organs occupy a distinct intensity band. In natural images, by contrast, dark pixels often can be the foreground object itself (tires, monitors), and intensity is weakly diagnostic of class. These priors, together with

the anatomical consistency, give this domain a natural advantage in learning from unlabeled data. Clinicians themselves rely on the same prior: an unfamiliar structure is identified by matching it against familiar ones, not interpreted from scratch.

Jointly, these facts describe an objective in which unlabeled and labeled supervision differ in kind, not merely in amount. We therefore decompose SSS into two responsibilities: *semantic acquisition*, which forms semantics of a class from its labeled anchors, and *semantic maintenance*, which consolidates the semantics already formed. Confidence reflects the extent of the model's semantic understanding, but it does not guarantee that this understanding is correct. The confidence-gated unlabeled branch cannot initiate semantics (F2), although it demonstrably maintains them (F3); acquisition must therefore be realized through training on labeled data. We evaluate the understanding the model believes it has: once this self-assessed understanding verifiably attains an expected level of correctness, the class is declared MASTERED. TriggerMatch accordingly treats supervision as a class-wise, time-varying allocation: while a class belongs to the acquisition set $\mathcal{C}_{\text{acq}}^{(t)}$, its positive anchors supervise learning; once MASTERED, it moves to the maintenance set $\mathcal{C}_{\text{maint}}^{(t)}$, where its semantics are rehearsed by the unlabeled branch rather than re-anchored.

The core idea of TriggerMatch is a class-wise routing of supervision: classes in the acquisition set $\mathcal{C}_{\text{acq}}^{(t)}$ are trained on their labeled anchors as usual, whereas for classes that have entered the maintenance set $\mathcal{C}_{\text{maint}}^{(t)}$ the corresponding gradient contributions are omitted from the labeled loss, i.e., these classes are masked. The semantics of the masked classes are instead maintained by the unlabeled branch. Labeled training is thereby dedicated to acquiring semantic understanding, and unlabeled training to maintaining it. Masking further balances semantic acquisition across classes on imbalanced datasets: it reduces the dilution of the acquisition gradients by mastered semantics, allowing the model to better understand such datasets.

**Mastery-Routed Labeled Objective.** Specifically, we recast the computation of the labeled CE and Dice losses so that both accumulate only over the voxels of acquisition-set classes:

$$\mathcal{L}_{\text{CE}}^{l} = -\frac{\sum_{v\in\Omega_l} \mathbb{I}\left[y_v^l \in \mathcal{C}_{\text{acq}}^{(t)}\right] \log p_{v,y_v^l}^l}{\max\left(\sum_{v\in\Omega_l} \mathbb{I}\left[y_v^l \in \mathcal{C}_{\text{acq}}^{(t)}\right], 1\right)}, \tag{4}$$

$$\mathcal{L}_{\text{Dice}}^{l} = \frac{1}{\max\left(\mathcal{C}_{\text{acq}}^{(t)}, 1\right)} \sum_{c\in\mathcal{C}_{\text{acq}}^{(t)}} \left(1 - \frac{2\sum_{v\in\Omega_l} \mathbb{I}\left[y_v^l = c\right] p_{v,c}^l}{\sum_{v\in\Omega_l} \mathbb{I}[y_v^l = c]\left(p_{v,c}^l + 1\right)}\right), \tag{5}$$

where $\Omega_l$ is the voxel set of the labeled batch. Only the classes in the maintenance set $\mathcal{C}_{\text{maint}}^{(t)}$ are retired from the labeled supervision, while the unlabeled branch of Eq. (3) remains unmasked, as required by Section 3.3.

### 3.3 The Role of Unlabeled Data: Maintenance, not Acquisition

Learning from unlabeled data depends strongly on what learning from labeled data has already established. If labeled training has not endowed the model with a sufficient understanding of the data, unlabeled data can hardly contribute beneficially to the overall training. As Figure 2 shows, when the labeled branch is detached from the training early on, the best weights that training on unlabeled data alone can finally attain hardly surpass the understanding that sufficient labeled training would have established. As Figure 3 shows, when training relies on gradients from unlabeled data alone, the model does not understand better; it fails even to maintain its existing understanding and instead degrades as training proceeds. Several factors may explain this degradation. One plausible cause is that supervision is still gated by high-confidence: with only limited high-confidence coverage, the model may become biased toward the background, and the resulting gradients crowd out its established understanding of foreground classes, so that overall semantic understanding deteriorates. This observation also indicates that the maintenance role of unlabeled data depends on the correction supplied by labeled data: high-confidence predictions on unlabeled data reflect only what the model already understands (F2), not the full content of a semantic class that remains to be learned. We therefore introduce a recovery mechanism for masked classes.

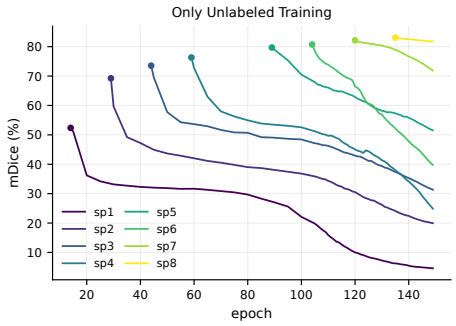



Figure 3: Validation mDice curves of runs in which labeled training is disabled at different stages ($sp$) and only training with unlabeled data.

Unlabeled data is exploited through two concrete mechanisms: pseudo-label supervision, and a consistency loss that compels the model to capture spatially complex information such as texture.

Taken together, the results indicate that unlabeled data cannot assume the responsibility of semantic acquisition; it can support semantic maintenance only under a restrictive condition, which our experiments identify concretely: once the corresponding class has attained a state that we term MASTERED, the model can temporarily rely on unlabeled data to maintain its understanding of that class. Handing acquisition responsibility to it early destabilizes training, whereas the labeled-only variant remains stable. This situation motivates both the warm-up and the strict entrance threshold.

### 3.4 The Components of TriggerMatch Design

We first define how the correctness of the model's understanding is evaluated, measured by the precision over the voxels that the model predicts with high-confidence. At each iteration $t$, the teacher predicts the labeled data with gradients disabled, and for every foreground class $c$ we form the high-confidence evaluator:

$$E_{c,t} = \frac{\sum_{v\in\Omega_l} \mathbb{I}\big[y_v^l = c \ \wedge\ \hat{y}_v^l = c \ \wedge\ s_v^l \geq \tau_a\big]}{\sum_{v\in\Omega_l} \mathbb{I}[y_v^l = c]}, \tag{6}$$

where $\Omega_l$ denotes the set of voxels in the labeled batch, and $\hat{y}_v^l$ and $s_v^l$ are the teacher's prediction and confidence on the labeled view. Where $\tau_a$ is the high-confidence filtering threshold, set to the same value as the threshold $\tau_u$ that gates the unlabeled loss, and the filter thus retains the voxels whose semantics the model already understands. The condition for a specific class to enter the maintenance set $\mathcal{C}_{\text{maint}}^{(t)}$ requires that the model has already achieved MASTERED of that semantic class. Because training draws a random crop of a fixed shape at every step, the evidence a class presents within the labeled batch varies from observation to observation. Each observation is therefore first typed by evidence: NO-GT, NO-PREDICTION, or EVALUABLE. NO-GT marks a class that is absent from the minibatch. NO-PREDICTION indicates no confident prediction, such an observation is explicit collapse evidence and is assigned $E_{c,t} = 0$. Meanwhile, we smooth $E_{c,t}$ with an EMA for stability, and an observation enters the update if and only if the class is EVALUABLE:

$$R_c^{(k)} = \beta R_c^{(k-1)} + (1-\beta)\, E_c^{(k)}, \tag{7}$$

When $R^{(k)} \geq \rho_{\text{in}}$ holds for $K$ consecutive iterations, we assign the class to the maintenance set $\mathcal{C}_{\text{maint}}^{(t)}$, i.e., we mark it as MASTERED. Where $k$ indexes the class's own evidence clock, so rare anatomy accumulates history at its own pace rather than being diluted by frequent organs. The partition of Eq. (4) is read off this state:

$$\mathcal{C}_{\text{acq}}^{(t)} = \{c : z_{c,t} = 0\}, \mathcal{C}_{\text{maint}}^{(t)} = \{c : z_{c,t} = 1\}, \tag{8}$$

**Hysteretic Transition and Recovery.** After a warm-up of $T_w$ epochs, an active class enters $\mathcal{C}_{\text{maint}}$ once $R_c \geq \rho_{\text{in}}$ on $K$ consecutive EVALUABLE observations; a mastered class returns to $\mathcal{C}_{\text{acq}}$ when its recall falls below $\rho_{\text{out}}$ on $K$ consecutive observations, or immediately upon $K$ consecutive NO-PREDICTION observations with all classes initialized active, $z_{c,0} = 0$:

$$z_{c,t+1} = \begin{cases} 1, & T \geq T_w,\ \nu_{c,t}^{\text{in}} \geq K, \\ 0, & \nu_{c,t}^{\text{out}} \geq K \ \vee\ \nu_{c,t}^{\text{np}} \geq K, \\ z_{c,t}, & \text{otherwise}, \end{cases} \tag{9}$$

Table 1: Comparison with SOTA semi-supervised methods on AMOS2022, FLARE2022 and Synapse under different label budgets. All entries are mean±std over 3 random seeds. mDice (%) ↑ and HD95 (mm) ↓ are reported. **Red bold**: best; **blue bold**: second best.

| Method | | AMOS2022 (volumes) | | | FLARE2022 (volumes) | | | Synapse (volumes) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 80 | 120 | 240 | 5 | 10 | 20 | 1 | 2 | 4 |
| SupOnly | mDice | 82.79±0.33 | 83.52±0.02 | 84.77±0.12 | 74.75±0.61 | 79.14±0.20 | 84.73±0.23 | 35.00±0.78 | 52.26±0.67 | 66.48±0.51 |
| | HD95 | 7.31±0.77 | 6.56±0.33 | 6.02±0.15 | 16.61±1.56 | 11.71±1.15 | 7.72±0.36 | 66.53±6.01 | 26.63±4.25 | 19.90±1.77 |
| FixMatch | mDice | 83.12±0.16 | 83.98±0.09 | 85.17±0.12 | 79.46±0.83 | 82.91±0.61 | 86.56±0.26 | 41.37±1.06 | 56.08±1.23 | 69.14±0.31 |
| | HD95 | 7.44±0.33 | 6.59±0.36 | 5.85±0.07 | 14.09±3.65 | 11.06±0.54 | 6.71±0.14 | **49.41±5.23** | 23.11±1.08 | 14.35±1.75 |
| FreeMatch | mDice | 83.34±0.32 | 83.87±0.38 | 85.17±0.22 | 70.06±1.28 | 79.70±0.54 | 86.24±0.14 | 22.24±3.35 | 49.45±6.46 | 65.66±1.50 |
| | HD95 | 7.84±0.57 | 6.57±0.40 | 6.04±0.18 | 48.88±3.39 | 27.18±7.08 | 8.60±1.30 | 111.65±16.30 | 56.07±21.08 | 35.00±0.89 |
| SoftMatch | mDice | 82.95±0.14 | 83.78±0.26 | 84.93±0.10 | 76.84±0.04 | 81.51±0.46 | 86.05±0.16 | 37.55±1.42 | 55.00±1.92 | 68.30±1.37 |
| | HD95 | 7.91±0.54 | 6.30±0.37 | 5.97±0.02 | 16.06±0.25 | 14.29±0.71 | 6.91±0.57 | 70.99±2.88 | 30.49±3.70 | 18.57±0.65 |
| CPS | mDice | 83.55±0.15 | 83.89±0.09 | 84.99±0.12 | 79.72±1.24 | 83.05±0.46 | 86.39±0.07 | 40.57±0.48 | **57.71±0.25** | **70.48±0.55** |
| | HD95 | 7.08±0.55 | **6.29±0.35** | 5.89±0.12 | 10.80±1.22 | **9.59±0.14** | **6.55±0.11** | **48.46±2.14** | **20.77±0.91** | **12.78±1.45** |
| ReFixMatch | mDice | **83.91±0.26** | **84.27±0.23** | **85.24±0.13** | **80.48±0.44** | **83.68±0.36** | 86.49±0.24 | **41.79±1.43** | 57.39±0.07 | 69.71±1.89 |
| | HD95 | **6.82±0.27** | 6.33±0.26 | 5.98±0.20 | 13.84±2.00 | 9.83±1.38 | **6.68±0.24** | 58.21±11.03 | 26.90±1.09 | 15.38±2.63 |
| SAFixMatch | mDice | 83.24±0.24 | 83.92±0.26 | 85.03±0.16 | 80.08±0.46 | 83.17±0.23 | **86.82±0.49** | 39.48±0.68 | 56.61±0.83 | 69.58±0.33 |
| | HD95 | 7.16±0.12 | 6.47±0.35 | **5.83±0.42** | **10.41±0.33** | 11.25±0.10 | 6.93±1.06 | 59.78±2.03 | 21.44±0.57 | 13.88±1.44 |
| GAMT | mDice | 81.72±0.51 | 81.91±0.68 | 83.40±0.15 | 65.29±3.11 | 67.42±6.64 | 85.69±0.14 | 17.73±0.77 | 46.65±7.42 | 65.82±3.19 |
| | HD95 | 7.83±0.40 | 7.23±0.30 | 6.59±0.27 | 77.79±5.98 | 74.37±24.68 | 11.43±0.51 | 144.46±4.27 | 87.38±23.10 | 32.47±15.45 |
| Ours | mDice | **84.65±0.22** | **85.61±0.08** | **86.40±0.18** | **82.63±0.54** | **84.76±0.19** | **87.21±0.34** | **47.99±2.67** | **62.07±0.16** | **72.94±0.93** |
| | HD95 | **6.72±0.66** | **5.90±0.14** | **5.40±0.12** | **10.74±1.37** | **8.65±0.15** | 6.71±0.40 | 50.94±7.74 | **19.20±2.63** | **13.59±2.27** |

Table 2: The parameter studies of $\tau_u$, $K$, and $T_w$ are all conducted on FLARE2022 under the 5-label setting. **Bold** marks the default setting.

| | $\tau_u$ | | | $K$ (iterations) | | | | $T_w$ (epochs) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.90 | **0.95** | 0.97 | 0 | 2 | **5** | 10 | 0 | 2 | **5** | 10 |
| mDice (%) ↑ | 81.90 | **82.83** | 82.43 | 82.17 | 82.04 | **82.83** | 81.62 | 81.40 | 81.36 | **82.83** | 82.05 |
| HD95 (mm) ↓ | 10.75 | **8.96** | 8.99 | 13.15 | 9.37 | **8.96** | 10.77 | 10.85 | 14.11 | **8.96** | 10.57 |

where $\nu^{\text{in}}_{c,t}$, $\nu^{\text{out}}_{c,t}$, and $\nu^{\text{np}}_{c,t}$ are streak counters on the evidence clock of class $c$, tallying the consecutive observations that satisfy the entry condition (EVALUABLE with $R_c \geq \rho_{\text{in}}$), the decline condition (EVALUABLE with $R_c \leq \rho_{\text{out}}$), and the collapse condition (NO-PREDICTION), respectively. Each counter resets to zero at the first observation that violates its condition, while NO-GT observations leave every counter unchanged.

# 4 Experiments

## 4.1 Datasets and Settings

**Datasets.** We evaluate on three benchmarks, Synapse, FLARE2022, and AMOS2022; per-benchmark details appear in Appendix B. Every experiment is run with three fixed random seeds, reporting the mean and standard deviation. Additional experimental details are provided in Appendix C.

**Implementation Details.** To ensure a fair comparison, we adopt the standard 3D U-Net as the unified backbone architecture for all image segmentation experiments. We implement all experiments using PyTorch and conduct all experiments on a single NVIDIA RTX 4090 GPU. The network is optimized using AdamW with a polynomial learning rate schedule. In our experiments, this corresponds to a total batch size of 8. Additional implementation details, including hyperparameters and dynamic updating rules, are provided in Appendix C.

**Compared Methods.** We compare TriggerMatch with both established baselines and recent state-of-the-art SSS approaches, including FixMatch (Sohn et al., 2020), GALoss with MT (Qi et al., 2024), CPS (Chen et al., 2021), SoftMatch(Chen et al., 2023), FreeMatch(Wang et al., 2023), SAFixMatch (Li et al., 2025) and ReFixMatch(Nguyen & Yang, 2023). For fairness, all methods are

Table 3: The parameter study of $\rho_{\text{in}}$ is conducted on FLARE2022 under the 20-label setting, and that of $\rho_{\text{out}}$ under the 5-label setting. **bold** marks the default setting.

| | $\rho_{\text{in}}$ | | | $\rho_{\text{out}}$ | | | |
|---|---|---|---|---|---|---|---|
| | 0.90 | **0.95** | 0.98 | 0.80 | 0.83 | **0.85** | 0.87 |
| mDice (%) ↑ | 85.48 | **87.37** | 87.88 | 81.57 | 81.63 | **82.83** | 82.72 |
| HD95 (mm) ↓ | 10.44 | **9.65** | 9.06 | 10.82 | 10.90 | **8.96** | 9.59 |

evaluated under the same data preprocessing, backbone and dataset partition protocol, while the method-specific training settings follow the standard configurations reported in their original papers whenever applicable.

### 4.2 Segmentation Results

Table 1 shows that TriggerMatch outperforms the state-of-the-art methods compared on all three datasets. On Synapse, mean Dice improves by 4.3 points on average, and by 6.2 in the lowest-label regime (one labeled volume); on AMOS2022, the average gain is 1.1 points; and on FLARE2022, TriggerMatch again ranks first, leading on eight of the nine label budgets. HD95 follows the same pattern: TriggerMatch attains the lowest boundary error on all three AMOS2022 budgets and on six of the nine settings overall, and the remaining exceptions lie within one standard deviation of the best baseline. Region overlap and boundary accuracy thus improve jointly, indicating that the gains are not traded against contour precision. We attribute the consistent improvements to a more principled route to semantic acquisition: routing divides the labor between the two data pools more efficiently, letting labeled supervision form class semantics while unlabeled supervision consolidates them, so that the model builds its understanding of the data rather than fights for it. Per-class segmentation results (mDice and HD95) on the three datasets, together with qualitative visualizations, are presented in detail in Appendix C.

### 4.3 Ablation Study

**Mechanism Ablations.** TriggerMatch rests on the division of labor of Section 3.3: labeled supervision acquires class semantics, and unlabeled supervision can only maintain them. We test each half of this mechanism on FLARE2022 (Ma & Wang, 2023) under the 5-label setting, running every probe for the full schedule of Section 4.1.

Disabling one loss term of the static objective from epoch $e$ onward isolates what each data pool contributes in Figure 2 (a). With the unlabeled loss disabled (M1), the best Dice settles at the level of labeled training alone, regardless of $e$: unlabeled supervision supplies the static method's gain over SupOnly, yet pushes no class beyond what its labeled anchors support. With the labeled loss disabled (M2), the best Dice never exceeds what the baseline had reached by epoch $e$, and the curve degrades from that level as training proceeds (Figure 3): disabling early leaves the model near chance, disabling late retains most of its understanding. Unlabeled supervision thus preserves the understanding that labeled supervision has produced, but does not extend it.

Rerouting supervision after warm-up isolates the direction of the routing in Figure 2 (b). Concentrating unlabeled supervision on $\mathcal{C}^{(t)}_{\text{maint}}$, on par with full TriggerMatch: mastered classes need only rehearsal, and their labeled anchors are redundant. The inverse routing (S2), which points the unlabeled branch at $\mathcal{C}^{(t)}_{\text{acq}}$, below SupOnly: high-confidence pseudo-labels reflect what the model already understands (F2) and cannot anchor classes whose semantics are still being formed. The S1–S2 asymmetry holds the amount of supervision fixed and varies only its direction, locating the mechanism in the routing itself rather than in the quantity of training signal. Note that neither S1 nor S2 is equipped with the recovery mechanism of Section 3.4. In both probes, retirement is irreversible: once a class enters the maintenance set $\mathcal{C}^{(t)}_{\text{maint}}$, it never returns to the acquisition set $\mathcal{C}^{(t)}_{\text{acq}}$. Fewer classes therefore receive labeled supervision than under SupOnly, and S2 consequently degrades below the SupOnly level.

We first test the mechanism that motivates TriggerMatch (Section 3.3), then evaluate the effectiveness and robustness of its components (Section 3.4), all under the protocol of Section 4.

Table 4: Ablation of the mask form on FLARE2022 (Ma & Wang, 2023) under the 5-label setting. mDice (%) ↑ and HD95 (mm) ↓ are reported.

| Mask form | mDice (%) ↑ | HD95 (mm) ↓ |
|---|---|---|
| Baseline | 78.79 | 12.85 |
| Low confidence | 79.85 | 13.30 |
| High confidence | 80.91 | 10.92 |
| Random | 79.99 | 14.24 |
| Soft | 81.62 | 9.81 |
| All | 82.83 | 8.96 |

Table 5: Ablation of the Trigger on FLARE2022 (Ma & Wang, 2023) under the 5-label setting. SupOnly and FixMatch denote the supervised-only baseline and the semi-supervised baseline, respectively. mDice (%) ↑ and HD95 (mm) ↓ are reported.

| Setting | mDice ↑ | HD95 ↓ |
|---|---|---|
| SupOnly | 72.31 | 16.67 |
| SupOnly + Trigger | 74.81 | 14.35 |
| FixMatch | 78.79 | 12.85 |
| FixMatch + Trigger | 82.83 | 8.96 |

**Parameter Study.** We analyze the sensitivity to the high-confidence threshold $\tau_u$ in Table 2. Too low a threshold admits voxels that the model has not reliably understood, and the resulting noise corrupts the results. Too high a threshold passes too few voxels, so that classes are seldom masked and the unlabeled branch receives only a sparse supervision signal, which likewise degrades performance. We additionally sweep the number of consecutive iterations over which the MASTERED state is retained, i.e., the exit patience $K$, to assess whether the controller is sufficiently sensitive. If the patience is too long, a degraded class remains masked and the model may drift in some cases for lack of labeled correction. If it is too short, classes leave the MASTERED state almost immediately and masking rarely takes effect. An appropriate intermediate value of $K$ is therefore required. To keep the conditions for entering the acquisition set $\mathcal{C}^{(t)}_{\text{acq}}$ and the maintenance set $\mathcal{C}^{(t)}_{\text{maint}}$ equally strict, we set a single patience $K_{\text{in}} = K_{\text{out}}$ rather than separate values. We also sweep the number of initial warm-up epochs $T_w$. Too few warm-up epochs allow masking to begin before the model's semantic understanding is sufficiently established, biasing the best understanding the model ultimately attains, so a sufficient warm-up also stabilizes semantic understanding. Too many, however, leave too little time for masking, and the model spends a larger share of training in the static, FixMatch-like phase. We also sweep the threshold for entering the maintenance set, $\rho_{\text{in}}$, applied to the smoothed precision $G^{(k)}$ (Table 3). As with the high-confidence threshold, too low a value introduces noise: a class may be certified while the model's understanding of it is still insufficient and its semantics still require substantial acquisition. Too high a value rarely triggers the entrance condition, and the model degenerates into the static, FixMatch-like baseline.

**Mask Mode Ablations.** We ablate the form of the mask applied to the labeled supervision of maintenance-set classes (Table 4), comparing five modes besides the no-mask baseline. All applies a hard mask and ignores the supervision gradients of all classes in $\mathcal{C}^{(t)}_{\text{maint}}$. Low confidence masks only the voxels of maintenance-set classes whose confidence falls below the high-confidence threshold $\tau_u$; High confidence masks the complementary set. Random draws a binary mask uniformly at random over the ground-truth voxels and masks the entries equal to zero. Soft forms a weight matrix from $1 - p_v$, where $p_v$ denotes the predicted probability of the ground-truth class at voxel $v$, yielding a smooth mask that interpolates between the Low-confidence and High-confidence modes.

## 5 Conclusion

We re-examined semi-supervised segmentation as an allocation between semantic acquisition and semantic maintenance, and built TriggerMatch, which routes the labeled objective by an online class-wise mastery evaluation with hysteretic, reversible retirement and a class-complete unlabeled branch. TriggerMatch achieves state-of-the-art performance on three public benchmarks. Limitations: This idea may be difficult to transfer directly to natural images: there, the diversity of data augmentation naturally lets consistency regularization attain a stronger understanding of the data, much as self-supervised and unsupervised learning enable a model to acquire part of the semantics without labels. The idea may likewise be difficult to transfer to inherently atypical medical images such as tumors: it presupposes a natural similarity prior between the semantic distribution of the labeled data and that of the unlabeled data, whereas some tumors exhibit highly irregular distributions that a few labeled examples can hardly cover.

### AI USE STATEMENT

In preparing this manuscript, the authors used large language model (LLM) tools to assist with drafting and polishing the text, and with language checking. No generative AI was used for code development, experiment execution, data analysis, or the mathematical content of the paper. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the content of the submission.

### REPRODUCIBILITY STATEMENT

All results are produced from a single codebase sharing one training harness; the controller hyperparameters (Table 14) are fixed across all nine dataset and label-budget settings. The benchmarks (AMOS2022, FLARE2022, and Synapse) are publicly released and used under their respective licenses; dataset statistics and data splits are described in Appendix B, and all training and evaluation configurations, including those of the baselines, in Appendix C. Every entry in Table 1 is the mean and standard deviation over three random seeds. All experiments run on a single [GPU model, e.g., NVIDIA 4090 (24GB)] GPU; source code, configuration files, split files, and per-run logs, including the controller trajectories underlying Section 4.2, will be released upon publication.

### ETHICS STATEMENT

This work uses publicly released, de-identified benchmark datasets (AMOS2022, FLARE2022, and Synapse) in accordance with their respective licenses and raises no human-subject or privacy concerns. The method is developed for research purposes only; any clinical application would require further validation on prospective, multi-center data.

## REFERENCES

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622. IEEE, 2021.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9538–9547, 2023.

Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219, 2015.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2017.

BA Landman, Z Xu, JE Iglesias, M Styner, TR Langerak, and A Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, 2015.

Jingyang Li, Jiachun Pan, Vincent Y. F. Tan, Kim-Chuan Toh, and Pan Zhou. Towards understanding why fixmatch generalizes better than supervised learning. In *International Conference on Learning Representations*, 2025.

Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *International conference on medical image computing and computer-assisted intervention*, pp. 552–561. Springer, 2020.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8801–8809, 2021a.

Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International conference on medical image computing and computer-assisted intervention*, pp. 318–329. Springer, 2021b.

Jun Ma and Bo Wang. *Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation: MICCAI 2022 Challenge, FLARE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*. Springer Nature, 2023.

Geoffrey J. McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975.

Khanh-Binh Nguyen and Joon-Sung Yang. Boosting semi-supervised learning by bridging high and low-confidence predictions. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1020–1030. IEEE, 2023.

Samet Oymak and Talha Cihad Gülcü. A theoretical characterization of semi-supervised learning with self-training for gaussian mixture models. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3601–3609, 2021.

Wenbo Qi, Jiafei Wu, and S. C. Chan. Gradient-aware for class-imbalanced semi-supervised medical image segmentation. In *Computer Vision – ECCV 2024*, volume 15113 of *Lecture Notes in Computer Science*, pp. 473–490, Cham, 2024. Springer Nature Switzerland.

Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22873–22882, 2024.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 605–613. Springer, 2019.

## A Augmentation Details

This appendix specifies the augmentation pipeline behind the views $a_w(x^u)$ and $a_s(x^u)$ introduced in Section 3.1. The pipeline is held fixed across all datasets and is shared verbatim by TriggerMatch and every method it is compared with, so that measured differences are attributable to how supervision is routed rather than to the augmentation distribution.

**Shared Spatial Transform.** Each training volume, labeled or unlabeled, is first reduced to a patch of size $S$ by a randomly placed crop $c$ ($S$=64×128×128 on all benchmarks of Section 4). Volumes shorter than $S$ along an axis are symmetrically zero-padded before cropping; in the unlabeled stream the voxels introduced by padding carry an ignore index and are removed from the set $\mathcal{V}_u$ entering Eq. (3), so the unsupervised objective is never evaluated on content outside the imaged field of view. The two views of an unlabeled volume derive from a *single* realization of the crop,

$$a_w(x^u) = c(x^u), \qquad a_s(x^u) = \tau_{\text{cut}} \circ \tau_{\text{int}} \circ c(x^u), \tag{10}$$

where $\tau_{\text{int}}$ and $\tau_{\text{cut}}$ denote the composite intensity and occlusion perturbations defined below. The weak view is thus the cropped patch itself, and both views remain in exact voxel-wise correspondence. No geometric transformation distinguishes the two views.

**Strong Perturbation.** $\tau_{\text{int}}$ chains three intensity operators in a fixed order, applied at the granularity of an individual volume line (one row of one axial slice), a convention inherited from per-channel volumetric augmentation (Isensee et al., 2021); $\tau_{\text{cut}}$ then applies 3D CutOut (DeVries & Taylor, 2017). Table 6 summarizes scope, activation probability, and parameters. For an activated line with mean $\mu$ and range $[m, M]$, the contrast operator applies $x' = \mu + \gamma\,(x - \mu)$ and clips the result to $[m, M]$; the gamma operator applies a power law to the range-normalized intensities and then renormalizes the line to its original mean and standard deviation, so that first- and second-order moments are preserved while the shape of the intensity histogram is perturbed. Factor exponents are drawn equiprobably from $\mathcal{U}(0.5, 1)$ or $\mathcal{U}(1, 1.5)$, symmetric in log scale about the identity.

Table 6: Strong-view perturbations, applied sequentially to the cropped patch. Intensity operators act per volume line within axial slices; for $S = 64\times128\times128$ the largest CutOut cuboid is $3\times6\times6$ voxels.

| Operator | Activation | Parameter | Effect |
|---|---|---|---|
| Brightness | slice w.p. 0.5; all its lines | $b \sim \mathcal{U}(0.5, 1.5)$ per line | $x' = b\,x$ |
| Contrast | volume line w.p. 0.5 | $\gamma \sim \mathcal{U}(0.5, 1) \vee \mathcal{U}(1, 1.5)$ | $x' = \mu + \gamma(x - \mu)$ |
| Gamma | volume line w.p. 0.5 | $\gamma$ as above | power law, moments retained |
| 3D CutOut | volume w.p. 0.7 | 5–15 cuboids; extents $\leq \lfloor d_i/20 \rfloor$ per axis | zero fill |

**Labeled Stream and Inference.** Labeled volumes receive the shared random crop and no further perturbation. The supervised branch therefore trains on the same intensity distribution that the weak view presents to the teacher. Validation and testing operate on full volumes with sliding-window inference and no augmentation.

## B Datasets and Experimental Protocol

**Benchmarks.** We evaluate on three public abdominal multi-organ segmentation benchmarks whose label spaces overlap extensively: FLARE2022 (Ma & Wang, 2023), AMOS2022 (Ji et al.,

2022), and Synapse (Landman et al., 2015). The labeled, validation, and test splits contain 20, 15, and 65 cases on FLARE2022, and 240, 30, and 30 cases on AMOS2022, respectively. Table 7 summarizes the splits and label budgets. Semi-supervised budgets are expressed as the number of labeled training volumes; all remaining training volumes of a benchmark are used without labels.

Table 7: Dataset statistics. $C$ denotes the number of foreground classes. Synapse (Landman et al., 2015) admits no separate validation set under the standard 18/12 patient split.

| Benchmark | $C$ | Labeled budgets | Unlabeled | Val. | Test | Epochs |
|---|---|---|---|---|---|---|
| FLARE2022 (Ma & Wang, 2023) | 14 | 5 / 10 / 20 | 2,000 | 15 | 80 | 75 |
| AMOS2022 (Ji et al., 2022) | 16 | 80 / 120 / 240 | 1,200 | 30 | 30 | 150 |
| Synapse (Landman et al., 2015) | 14 | 1 / 2 / 4 | 17 / 16 / 14 | — | 12 | 3,000 |

The three label spaces are refinements of a common core of twelve organs: liver, spleen, left and right kidney, pancreas, stomach, esophagus, gallbladder, aorta, inferior vena cava, and the left and right adrenal glands. FLARE2022 adds the duodenum; AMOS2022 adds the duodenum, the bladder, and the prostate or uterus; Synapse instead adds the portal and splenic vein. This near-nested overlap is what makes the cross-dataset anatomical regularity discussed in the Introduction meaningful, and it keeps the class imbalance structure—from voluminous organs such as the liver to sub-centimeter adrenal glands—comparable across benchmarks.

**Preprocessing and training protocol.** All volumes are resampled during preprocessing to a common anisotropic spacing of $(2.5, 1.25, 1.25)$ mm (FLARE2022 (Ma & Wang, 2023), AMOS2022) and cached; intensities are consumed at the preprocessed scale, and the augmentation of Appendix A operates on this representation. Every method compared in this paper—TriggerMatch and all baselines—shares the training loop of Section 3.1: the same `3D-UNet`-style 3D architecture, AdamW with initial learning rate $0.1$ and weight decay $10^{-3}$ under a polynomial decay schedule of exponent $0.9$, the pseudo-label threshold $\tau_u = 0.95$, and the teacher update of Eq. (1) with $\alpha_{\max} = 0.996$. Each optimization step consumes $B_l$ labeled and $B_l$ unlabeled crops of size $64\times128\times128$, with $B_l$=4 on FLARE2022 (Ma & Wang, 2023) and AMOS2022. One epoch enumerates the labeled stream while unlabeled volumes are resampled uniformly, so the iteration budget grows with the labeled budget (e.g., 375 steps at 20 labels versus 750 at 40 on FLARE2022 (Ma & Wang, 2023)); on Synapse the labeled batch comprises the entire labeled set ($B_l \in \{1, 2, 4\}$), which fixes the budget at 3,000 steps across all label regimes. All results are averaged over three random seeds.

**Evaluation.** Predictions are produced by sliding-window inference over the full volume, with windows of $64\times128\times128$ (FLARE2022, Synapse) or $64\times160\times160$ (AMOS2022). We report the mean Dice score over the $C$ foreground classes. On FLARE2022 (Ma & Wang, 2023) and AMOS2022 the checkpoint with the highest validation Dice is reported; on Synapse, which reserves no held-out validation set, all methods are trained for the fixed budget of Table 7 and evaluated with the final checkpoint, so that no held-out data participates in model selection.

## C Experiments Details

Figures 4, 6 and 5 present qualitative comparisons of the segmentation results, both from multiple viewpoints and as a whole. Tables 8 and 9 (AMOS2022), 10 and 11 (FLARE2022 (Ma & Wang, 2023)), and 12 and 13 (Synapse (Landman et al., 2015)) report the per-class mDice and HD95 values on the three datasets, respectively.

## D TriggerMatch Details

**Hyperparameters and Overhead.** Table 14 lists the controller hyperparameters in two layers, a measurement layer that reads out coverage and a decision layer that acts on it. All values, including the unlabeled weight $\lambda_u$=1.0, are held fixed across all nine dataset and label-budget settings; only crop size, learning rate, iteration count, and $\tau_u$ follow the per-dataset configuration shared by every method. The controller costs one additional teacher forward pass on the labeled minibatch per

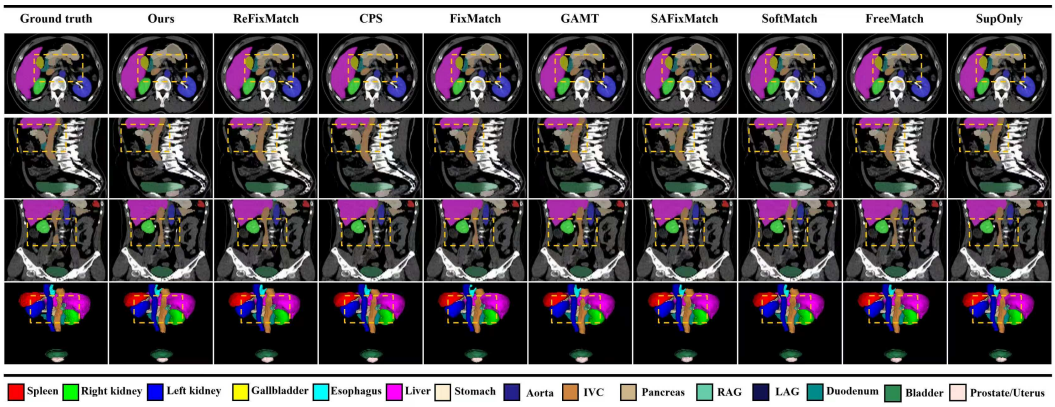



Figure 4: Qualitative comparison of different semi-supervised segmentation methods on the AMOS2022 dataset.

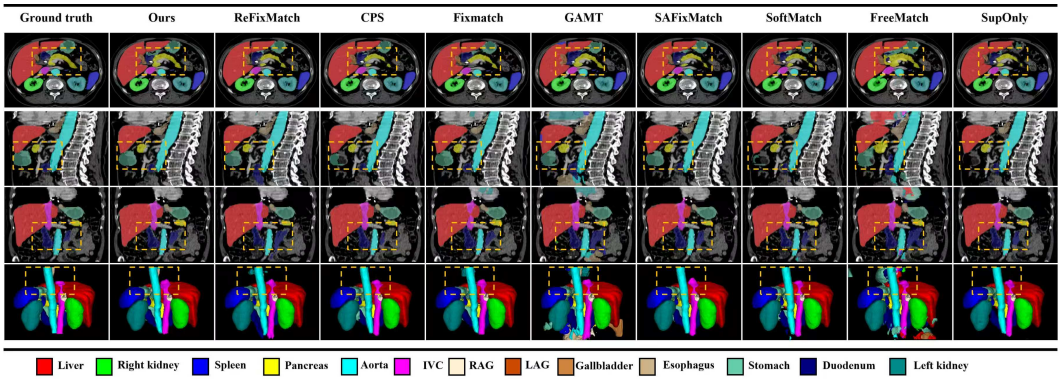



Figure 5: Qualitative comparison of different semi-supervised segmentation methods on the FLARE2022 (Ma & Wang, 2023) dataset.

iteration and constant bookkeeping; inference and the deployed model are identical to the static baseline. Figure 7 illustrates the overall framework.

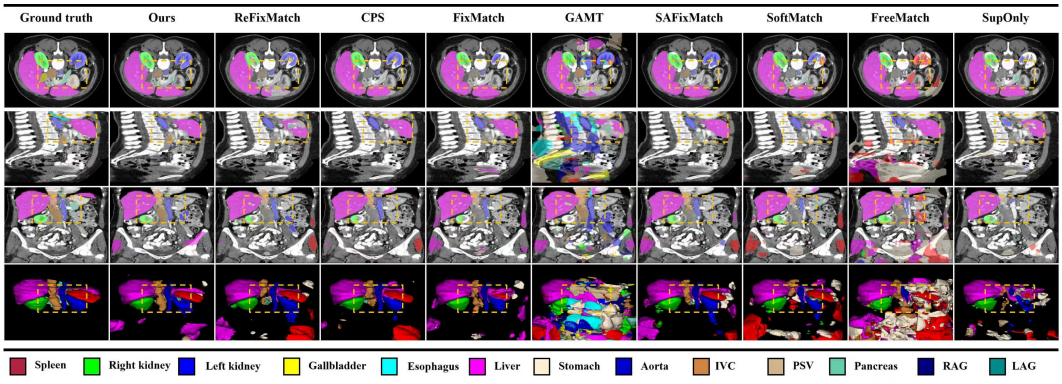



Figure 6: Qualitative comparison of different semi-supervised segmentation methods on the Synapse (Landman et al., 2015) dataset.

Table 8: Per-organ Dice (%)↑ on AMOS with 80, 120, 240 labeled cases. Per-organ values are averaged over seeds (1024, 2026, 2027); Avg. Dice is mean±std over the same seeds. **Red bold**: best; **blue bold**: second best (among semi-supervised methods).

| Method | Spl | R-Kid | L-Kid | Gall | Eso | Liv | Sto | Aor | IVC | Pan | RAG | LAG | Duo | Bla | P/U | Avg. Dice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **80 Labeled Cases** | | | | | | | | | | | | | | | | |
| SupOnly | 95.9 | 95.6 | 92.3 | 77.8 | 76.5 | 97.2 | 89.3 | 93.7 | 87.0 | 78.2 | 68.0 | 68.1 | 71.1 | 82.0 | 69.1 | 82.79±0.33 |
| FixMatch (Sohn et al., 2020) | 95.9 | 95.7 | 92.6 | 77.3 | 76.9 | 97.4 | 88.8 | 93.9 | 87.4 | 78.8 | 68.2 | 68.1 | 71.9 | 83.0 | 71.0 | 83.12±0.16 |
| FreeMatch (Wang et al., 2023) | 95.9 | 95.5 | 92.5 | 76.3 | 76.9 | 97.2 | 88.6 | 93.8 | **87.5** | 79.5 | 69.2 | 69.3 | 72.9 | 83.4 | 71.5 | 83.34±0.32 |
| SoftMatch (Chen et al., 2023) | 95.8 | 95.7 | 92.4 | **79.8** | 76.9 | 97.1 | 88.2 | 93.8 | 87.4 | 78.6 | 69.8 | 67.5 | 71.8 | 82.4 | 66.9 | 82.95±0.14 |
| CPS (Chen et al., 2021) | **96.1** | **95.7** | **92.7** | 78.7 | 76.5 | 97.4 | **89.4** | **94.1** | 87.5 | 78.9 | **70.8** | 69.6 | 73.1 | 82.6 | 70.2 | 83.55±0.15 |
| ReFixMatch (Nguyen & Yang, 2023) | **96.2** | 95.7 | **92.8** | 78.6 | **77.4** | **97.4** | **89.4** | **94.0** | 87.4 | **80.1** | 68.0 | **70.6** | **73.2** | **84.3** | **73.7** | **83.91±0.26** |
| SAFixMatch (Li et al., 2025) | 96.0 | **95.8** | 92.6 | 77.7 | 76.2 | **97.4** | 89.4 | 93.9 | 87.4 | 79.4 | 68.7 | 69.7 | 72.4 | 83.7 | 68.2 | 83.24±0.24 |
| GAMT (Qi et al., 2024) | 95.1 | 94.2 | 91.2 | 76.2 | 74.9 | 96.8 | 88.7 | 91.8 | 85.1 | 79.1 | 61.0 | 62.3 | 72.0 | **84.6** | 72.9 | 81.72±0.51 |
| **TriggerMatch (Ours)** | 95.9 | 95.6 | 92.5 | **80.0** | **79.5** | 97.0 | 88.3 | 93.9 | **87.8** | **79.8** | **73.2** | **74.4** | **73.5** | 83.9 | **74.4** | **84.65±0.22** |
| **120 Labeled Cases** | | | | | | | | | | | | | | | | |
| SupOnly | 96.1 | 95.8 | 92.5 | 80.5 | 78.1 | 97.2 | 89.8 | 93.7 | 87.6 | 79.5 | 67.9 | 66.9 | 72.5 | 83.1 | 71.6 | 83.52±0.02 |
| FixMatch (Sohn et al., 2020) | **96.2** | 95.8 | **92.6** | 79.8 | 77.9 | 97.4 | **90.1** | 93.9 | 87.8 | 80.0 | 69.1 | **70.1** | 73.9 | 83.3 | 71.8 | 83.98±0.09 |
| FreeMatch (Wang et al., 2023) | 95.9 | 95.4 | 92.4 | 79.5 | 76.5 | 97.3 | 89.3 | 93.8 | 87.6 | 79.8 | **69.9** | 69.7 | 73.6 | 82.9 | **74.5** | 83.87±0.38 |
| SoftMatch (Chen et al., 2023) | 96.2 | 95.7 | 92.5 | 79.2 | 78.1 | 97.4 | 89.8 | 94.0 | 87.8 | 79.6 | 69.1 | 68.5 | 73.4 | 83.7 | 71.9 | 83.78±0.26 |
| CPS (Chen et al., 2021) | 96.2 | **95.8** | 92.6 | 80.9 | 77.6 | 97.4 | 90.0 | **94.0** | 87.7 | 79.9 | 68.9 | 69.5 | 73.6 | 83.0 | 71.3 | 83.89±0.09 |
| ReFixMatch (Nguyen & Yang, 2023) | **96.2** | 95.7 | **92.6** | **81.1** | **78.5** | **97.4** | **90.3** | 93.8 | **88.0** | **80.5** | 69.6 | 69.3 | **74.8** | 83.7 | 72.6 | **84.27±0.23** |
| SAFixMatch (Li et al., 2025) | 96.1 | **95.8** | 92.5 | 79.6 | 77.5 | **97.4** | 90.0 | 94.0 | 87.8 | 79.6 | 69.5 | 69.3 | 73.3 | 82.7 | 73.5 | 83.92±0.26 |
| GAMT (Qi et al., 2024) | 95.1 | 94.6 | 91.4 | 80.6 | 73.4 | 97.1 | 89.9 | 91.6 | 85.8 | 79.6 | 61.6 | 58.2 | 72.9 | **83.8** | 73.0 | 81.91±0.68 |
| **TriggerMatch (Ours)** | 95.8 | 95.7 | 92.5 | **83.3** | **80.5** | 97.1 | 89.7 | **94.0** | **88.1** | **80.3** | **75.3** | **75.6** | **74.8** | **84.6** | **76.9** | **85.61±0.08** |
| **240 Labeled Cases** | | | | | | | | | | | | | | | | |
| SupOnly | 96.1 | 95.9 | 92.7 | 81.0 | 79.6 | 97.5 | 90.4 | 94.2 | 88.3 | 81.4 | 70.0 | 69.4 | 75.9 | 84.1 | 75.0 | 84.77±0.12 |
| FixMatch (Sohn et al., 2020) | **96.2** | **96.0** | **92.8** | 82.3 | 80.1 | **97.5** | 90.6 | **94.3** | **88.6** | 81.8 | 70.3 | 70.8 | 76.3 | 83.7 | 76.4 | 85.17±0.12 |
| FreeMatch (Wang et al., 2023) | 96.1 | 95.9 | 92.7 | 81.2 | 79.8 | 97.5 | **90.8** | 94.2 | 88.6 | 81.9 | 70.5 | 70.4 | 76.7 | 84.1 | **77.3** | 85.17±0.22 |
| SoftMatch (Chen et al., 2023) | 96.2 | **96.0** | **92.8** | 80.3 | 79.5 | 97.5 | 90.5 | 94.2 | 88.2 | 81.9 | 70.3 | 70.0 | 75.7 | **84.3** | 76.8 | 84.93±0.10 |
| CPS (Chen et al., 2021) | 96.0 | 95.9 | 92.7 | 81.0 | 79.9 | 97.5 | 90.5 | **94.2** | 88.4 | 81.6 | **71.2** | **70.9** | 76.3 | 83.6 | 74.9 | 84.99±0.12 |
| ReFixMatch (Nguyen & Yang, 2023) | **96.2** | 95.9 | 92.8 | 81.2 | **80.3** | 97.5 | 90.3 | 94.2 | 88.5 | **82.3** | 71.0 | 70.5 | **76.8** | 84.1 | 77.1 | **85.24±0.13** |
| SAFixMatch (Li et al., 2025) | 96.1 | 95.9 | 92.8 | 82.0 | 79.8 | **97.5** | **90.6** | 94.1 | 88.5 | 81.9 | 70.7 | 70.4 | 76.6 | 83.4 | 75.1 | 85.03±0.16 |
| GAMT (Qi et al., 2024) | 95.2 | 94.9 | 91.8 | **82.5** | 75.1 | 97.1 | 90.3 | 92.6 | 86.9 | 81.7 | 65.0 | 63.6 | 76.2 | 83.2 | 75.0 | 83.40±0.15 |
| **TriggerMatch (Ours)** | 95.8 | 95.7 | 92.6 | **84.7** | **82.5** | 97.2 | 89.5 | 94.2 | **88.6** | **82.5** | **75.5** | **77.3** | **77.4** | **84.5** | **77.9** | **86.40±0.18** |

Table 9: Per-organ HD95 (mm)↓ on AMOS with 80, 120, 240 labeled cases. Per-organ values are averaged over seeds (1024, 2026, 2027); Avg. HD95 is mean±std over the same seeds. **Red bold**: best; **blue bold**: second best (among semi-supervised methods).

| Method | Spl | R-Kid | L-Kid | Gall | Eso | Liv | Sto | Aor | IVC | Pan | RAG | LAG | Duo | Bla | P/U | Avg. HD95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **80 Labeled Cases** | | | | | | | | | | | | | | | | |
| SupOnly | 4.1 | 2.0 | 8.9 | 10.7 | 7.0 | 3.6 | 7.1 | 7.0 | 4.2 | 8.5 | 4.9 | 5.4 | 13.7 | 8.8 | 13.8 | 7.31±0.77 |
| FixMatch (Sohn et al., 2020) | 5.2 | 2.0 | 8.2 | **11.4** | 7.5 | **3.3** | 9.4 | 7.6 | 4.0 | 8.9 | 4.4 | 5.5 | 12.6 | 8.5 | 13.1 | 7.44±0.33 |
| FreeMatch (Wang et al., 2023) | 3.7 | 2.1 | **6.8** | 13.6 | 7.1 | 3.5 | 10.0 | 6.6 | 4.2 | 8.9 | 4.7 | 5.7 | 12.6 | 8.7 | 19.3 | 7.84±0.57 |
| SoftMatch (Chen et al., 2023) | 6.7 | 2.0 | 9.0 | 11.7 | 7.0 | 3.9 | 10.2 | 6.5 | **4.0** | 8.8 | 4.8 | 5.7 | 13.0 | 11.3 | 14.2 | 7.91±0.54 |
| CPS (Chen et al., 2021) | **2.3** | 2.0 | 7.3 | **10.0** | 8.3 | 3.5 | **8.6** | **6.2** | 4.0 | 9.3 | **4.3** | **5.4** | 12.1 | 8.4 | 14.5 | 7.08±0.55 |
| ReFixMatch (Nguyen & Yang, 2023) | **2.2** | **2.0** | **7.1** | 11.8 | 6.8 | **3.4** | 8.8 | 6.6 | 4.1 | 8.3 | 4.5 | 5.5 | **11.3** | 8.0 | **11.9** | **6.82±0.27** |
| SAFixMatch (Li et al., 2025) | 3.9 | **2.0** | 8.6 | 11.6 | 7.2 | 3.4 | **7.7** | 6.4 | 4.0 | 8.6 | 4.6 | 5.6 | 12.6 | 8.0 | 13.6 | 7.16±0.12 |
| GAMT (Qi et al., 2024) | 2.8 | 2.6 | 9.5 | 16.9 | **5.4** | 3.8 | 10.1 | 8.2 | 4.8 | **7.0** | 6.4 | 6.5 | 11.8 | **7.7** | 13.9 | 7.83±0.40 |
| **TriggerMatch (Ours)** | 3.8 | 2.1 | 7.4 | 12.2 | **5.4** | 3.8 | 9.0 | **5.9** | **4.0** | **7.4** | **4.0** | **4.7** | **11.1** | **7.4** | **12.8** | **6.72±0.66** |
| **120 Labeled Cases** | | | | | | | | | | | | | | | | |
| SupOnly | 2.2 | 2.0 | 7.3 | 10.9 | 5.7 | 3.6 | 6.0 | 6.2 | 4.0 | 7.5 | 4.6 | 5.8 | 10.9 | 8.6 | 13.1 | 6.56±0.33 |
| FixMatch (Sohn et al., 2020) | **2.2** | **2.0** | 7.2 | 12.2 | 5.5 | 3.4 | 6.6 | 6.1 | 4.0 | 7.8 | 4.4 | **5.0** | **10.6** | 8.7 | 13.1 | 6.59±0.36 |
| FreeMatch (Wang et al., 2023) | 2.4 | 2.1 | 4.6 | 11.8 | 5.8 | **3.3** | 8.3 | 6.3 | 4.1 | 7.6 | 4.7 | 5.5 | 11.2 | 8.8 | **11.9** | 6.57±0.40 |
| SoftMatch (Chen et al., 2023) | 2.2 | 2.0 | 4.6 | 11.4 | **5.1** | 3.5 | **6.0** | **5.7** | 3.9 | 7.7 | 4.6 | 5.6 | 10.7 | 8.5 | 13.0 | 6.30±0.37 |
| CPS (Chen et al., 2021) | 2.3 | 2.0 | **4.5** | 10.0 | 6.0 | 3.4 | **6.0** | 5.8 | 3.9 | 8.1 | **4.4** | 5.3 | 10.7 | 9.3 | 12.7 | **6.29±0.35** |
| ReFixMatch (Nguyen & Yang, 2023) | **2.1** | 2.0 | **4.5** | 11.4 | 5.2 | **3.3** | 6.8 | 6.5 | **3.8** | 7.6 | 4.5 | 5.5 | 10.7 | 8.7 | 12.4 | 6.33±0.26 |
| SAFixMatch (Li et al., 2025) | 2.2 | **2.0** | 8.6 | **8.8** | 6.0 | 3.3 | 7.0 | **5.6** | 4.0 | 7.7 | 4.4 | 5.2 | 11.1 | 8.5 | 12.6 | 6.47±0.35 |
| GAMT (Qi et al., 2024) | 2.6 | 2.5 | 9.1 | 11.7 | 6.2 | 3.7 | 7.7 | 8.6 | 4.5 | **6.8** | 6.1 | 7.2 | 11.6 | **7.8** | 12.3 | 7.23±0.30 |
| **TriggerMatch (Ours)** | 2.6 | 2.0 | 6.0 | **9.3** | **4.4** | 3.8 | 6.7 | 5.7 | **3.8** | **7.1** | **3.7** | **4.4** | **10.4** | **7.8** | **10.9** | **5.90±0.14** |
| **240 Labeled Cases** | | | | | | | | | | | | | | | | |
| SupOnly | 2.4 | 2.0 | 5.2 | 8.7 | 4.5 | 3.1 | 5.4 | 5.7 | 3.8 | 7.3 | 4.6 | 5.0 | 9.4 | 9.8 | 13.3 | 6.02±0.15 |
| FixMatch (Sohn et al., 2020) | 2.3 | **1.9** | 5.3 | **7.7** | 4.4 | 3.2 | 5.5 | 5.7 | **3.7** | 6.8 | 4.4 | **4.6** | 9.4 | 11.1 | 11.8 | 5.85±0.07 |
| FreeMatch (Wang et al., 2023) | **2.2** | 2.0 | 7.7 | 8.4 | 4.8 | 3.1 | **5.4** | 5.7 | 3.8 | 7.1 | 4.6 | 4.9 | 9.7 | 9.8 | **11.7** | 6.04±0.18 |
| SoftMatch (Chen et al., 2023) | 2.3 | **1.9** | **1.9** | 9.5 | 4.8 | 3.0 | 5.9 | **5.7** | 3.9 | 6.5 | 4.3 | 5.1 | 9.3 | 10.3 | 15.1 | 5.97±0.02 |
| CPS (Chen et al., 2021) | 2.6 | 2.0 | 3.5 | 8.1 | 4.5 | **3.0** | 6.0 | **5.6** | **3.8** | 6.6 | 4.1 | 4.9 | **8.9** | 10.3 | 14.6 | 5.89±0.12 |
| ReFixMatch (Nguyen & Yang, 2023) | **2.1** | 1.9 | 8.0 | 7.9 | **4.2** | 3.1 | **5.4** | 5.8 | 3.8 | **6.2** | **4.1** | 4.9 | 9.2 | **9.4** | 13.5 | 5.98±0.20 |
| SAFixMatch (Li et al., 2025) | 2.2 | 1.9 | 4.4 | 7.8 | 4.4 | **3.0** | 5.6 | 5.7 | 3.8 | 6.6 | 4.1 | 4.7 | 9.3 | 10.3 | 13.5 | **5.83±0.42** |
| GAMT (Qi et al., 2024) | 2.8 | 2.5 | 4.7 | 9.9 | 5.5 | 3.7 | 6.0 | 8.7 | 4.2 | **6.0** | 5.6 | 6.1 | 9.8 | 10.3 | 13.4 | 6.59±0.27 |
| **TriggerMatch (Ours)** | 2.6 | 2.1 | **3.4** | **7.2** | **4.2** | 3.4 | 5.8 | 5.8 | 3.8 | 6.2 | **3.5** | **3.8** | **8.7** | **8.8** | **11.7** | **5.40±0.12** |

Table 10: Per-organ Dice (%)↑ on FLARE2022 with 5, 10, 20 labeled cases. Per-organ values are averaged over seeds (1024, 2026, 2027); Avg. Dice is mean±std over the same seeds. **Red bold**: best; **blue bold**: second best (among semi-supervised methods).

| Method | Liv | R-Kid | Spl | Pan | Aor | IVC | RAG | LAG | Gall | Eso | Sto | Duo | L-Kid | Avg. Dice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **5 Labeled Cases** | | | | | | | | | | | | | | |
| SupOnly | 91.6 | 86.9 | 91.8 | 65.2 | 92.6 | 81.2 | 56.3 | 53.3 | 66.9 | 65.3 | 79.4 | 57.0 | 84.3 | 74.75±0.61 |
| FixMatch (Sohn et al., 2020) | 96.9 | **93.0** | 94.2 | **77.2** | 94.1 | **86.5** | 55.7 | 49.6 | 71.4 | 66.5 | 87.6 | 64.9 | **95.2** | 79.46±0.83 |
| FreeMatch (Wang et al., 2023) | 90.4 | 82.1 | 82.3 | 63.0 | 84.9 | 79.9 | 52.7 | 52.8 | 59.6 | 60.9 | 68.9 | 53.8 | 79.6 | 70.06±1.28 |
| SoftMatch (Chen et al., 2023) | 96.8 | 90.0 | 93.6 | 67.7 | 93.0 | 84.0 | 57.6 | 51.4 | **73.1** | 64.9 | 85.5 | 59.9 | 81.4 | 76.84±0.04 |
| CPS (Chen et al., 2021) | **97.5** | 92.2 | **95.8** | 73.7 | **94.4** | **86.5** | 57.2 | 53.4 | **72.0** | 69.1 | **89.2** | 62.9 | 92.3 | 79.72±1.24 |
| ReFixMatch (Nguyen & Yang, 2023) | 97.3 | 91.2 | 95.0 | 75.0 | 93.9 | 85.7 | **60.4** | **61.2** | 66.6 | **73.6** | 88.3 | 66.9 | 91.1 | **80.48±0.44** |
| SAFixMatch (Li et al., 2025) | **97.5** | **92.6** | **95.4** | 74.6 | 94.0 | 85.8 | 53.2 | 55.0 | 71.9 | 69.4 | **89.3** | **67.4** | **93.9** | 80.08±0.46 |
| GAMT (Qi et al., 2024) | 93.0 | 84.2 | 85.3 | 68.0 | 81.1 | 77.5 | 39.0 | 32.2 | 46.1 | 28.7 | 77.6 | 57.8 | 78.4 | 65.29±3.11 |
| **TriggerMatch (Ours)** | 96.3 | 90.9 | 94.7 | **77.5** | **94.2** | 84.5 | **70.6** | **72.9** | 68.0 | **76.0** | 87.2 | **69.4** | 92.0 | **82.63±0.54** |
| **10 Labeled Cases** | | | | | | | | | | | | | | |
| SupOnly | 94.6 | 89.1 | 93.3 | 71.8 | 93.9 | 84.1 | 66.3 | 61.0 | 69.2 | 72.9 | 86.4 | 63.2 | 83.1 | 79.14±0.20 |
| FixMatch (Sohn et al., 2020) | 96.9 | **92.3** | 94.8 | 78.0 | 94.3 | 85.4 | 74.9 | 69.5 | 66.1 | 77.7 | 89.5 | 69.6 | 88.9 | 82.91±0.61 |
| FreeMatch (Wang et al., 2023) | 93.8 | 83.9 | 92.8 | 74.3 | 92.3 | 84.9 | 69.8 | 67.9 | 69.6 | 77.2 | 81.8 | 68.2 | 79.6 | 79.70±0.54 |
| SoftMatch (Chen et al., 2023) | 95.5 | 88.0 | 94.1 | 76.4 | 94.0 | 84.5 | 72.5 | 70.9 | **70.2** | 77.9 | 86.5 | 67.2 | 82.0 | 81.51±0.46 |
| CPS (Chen et al., 2021) | **96.9** | **92.8** | 94.8 | 78.3 | **94.4** | **85.9** | **75.7** | **71.5** | 65.6 | 77.7 | 89.4 | 68.4 | 88.3 | 83.05±0.46 |
| ReFixMatch (Nguyen & Yang, 2023) | **97.1** | 91.8 | **94.9** | **79.3** | 93.6 | 85.8 | 73.8 | 71.2 | 68.4 | **79.0** | **90.0** | **70.5** | **92.4** | **83.68±0.36** |
| SAFixMatch (Li et al., 2025) | 96.3 | 90.7 | **95.6** | 78.3 | 94.3 | 85.8 | 74.7 | 71.2 | 66.7 | 78.2 | 89.5 | 69.2 | 90.8 | 83.17±0.23 |
| GAMT (Qi et al., 2024) | 92.7 | 84.7 | 87.3 | 64.8 | 80.4 | 78.7 | 41.9 | 30.3 | 45.2 | 47.4 | 83.0 | 59.3 | 80.6 | 67.42±6.64 |
| **TriggerMatch (Ours)** | 96.9 | 92.1 | 94.5 | **80.2** | **94.8** | **86.2** | **75.6** | **72.8** | **70.4** | **80.9** | **90.9** | **72.8** | **93.5** | **84.76±0.19** |
| **20 Labeled Cases** | | | | | | | | | | | | | | |
| SupOnly | 97.9 | 94.4 | 95.3 | 79.5 | 95.1 | 87.8 | 75.8 | 69.1 | 74.8 | 78.0 | 90.0 | 70.9 | 92.8 | 84.73±0.23 |
| FixMatch (Sohn et al., 2020) | **97.9** | 94.0 | **96.4** | 81.5 | 94.9 | 88.4 | 79.6 | 75.6 | **77.6** | 81.0 | 90.8 | 72.9 | 94.7 | 86.56±0.26 |
| FreeMatch (Wang et al., 2023) | 97.8 | 93.5 | 95.1 | 81.3 | 94.8 | 87.9 | 79.1 | 75.5 | 75.8 | **81.6** | 89.9 | 73.8 | 95.1 | 86.24±0.14 |
| SoftMatch (Chen et al., 2023) | 97.9 | 94.1 | 95.3 | 80.9 | 94.8 | 87.9 | 79.3 | 74.6 | 76.7 | 80.3 | 90.1 | 72.7 | 94.0 | 86.05±0.16 |
| CPS (Chen et al., 2021) | **98.0** | 94.1 | 95.4 | 81.8 | **95.2** | **88.7** | **80.3** | 74.9 | 75.6 | 81.0 | 90.2 | 73.1 | 94.8 | 86.39±0.07 |
| ReFixMatch (Nguyen & Yang, 2023) | 97.9 | 94.1 | 95.4 | 81.7 | 94.9 | 87.6 | 79.9 | **77.1** | 75.8 | 80.4 | 90.9 | 73.4 | 95.2 | 86.49±0.24 |
| SAFixMatch (Li et al., 2025) | 97.8 | **94.4** | **96.9** | 81.3 | 95.0 | 88.5 | **80.4** | 75.8 | **78.3** | 81.5 | 90.4 | 73.2 | **95.3** | **86.82±0.49** |
| GAMT (Qi et al., 2024) | 97.9 | 93.9 | 95.7 | **82.4** | 94.4 | 88.1 | 75.3 | 73.5 | 70.9 | 79.9 | **91.1** | **75.9** | 95.1 | 85.69±0.14 |
| **TriggerMatch (Ours)** | 97.7 | **94.3** | 95.1 | **83.1** | **95.4** | **88.7** | 79.1 | **76.9** | 76.9 | **82.7** | **92.1** | **76.1** | **95.7** | **87.21±0.34** |

Table 11: Per-organ HD95 (mm)↓ on FLARE2022 with 5, 10, 20 labeled cases. Per-organ values are averaged over seeds (1024, 2026, 2027); Avg. HD95 is mean±std over the same seeds. **Red bold**: best; **blue bold**: second best (among semi-supervised methods).

| Method | Liv | R-Kid | Spl | Pan | Aor | IVC | RAG | LAG | Gall | Eso | Sto | Duo | L-Kid | Avg. HD95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **5 Labeled Cases** | | | | | | | | | | | | | | |
| SupOnly | 15.0 | 23.3 | 31.4 | 13.4 | 4.7 | 13.0 | 10.2 | 12.3 | 22.2 | 11.4 | 21.0 | 20.6 | 17.4 | 16.61±1.56 |
| FixMatch (Sohn et al., 2020) | 13.0 | **12.4** | 23.4 | 12.6 | 6.0 | 10.6 | 11.2 | 12.9 | 23.9 | 13.3 | 14.6 | 19.7 | 9.5 | 14.09±3.65 |
| FreeMatch (Wang et al., 2023) | 63.3 | 64.1 | 88.3 | 39.0 | 16.8 | 13.7 | 27.1 | 20.7 | 72.9 | 23.9 | 85.1 | 53.9 | 66.6 | 48.88±3.39 |
| SoftMatch (Chen et al., 2023) | 10.8 | 18.6 | 15.4 | 12.9 | **3.7** | 8.3 | 11.2 | 15.7 | **21.7** | 11.7 | 39.1 | 17.7 | 22.0 | 16.06±0.25 |
| CPS (Chen et al., 2021) | **6.1** | 18.5 | **4.7** | **9.3** | **2.4** | **7.0** | 8.9 | 9.8 | 26.0 | 11.7 | **10.9** | 18.7 | **6.4** | 10.80±1.22 |
| ReFixMatch (Nguyen & Yang, 2023) | 6.1 | 20.6 | 8.1 | 9.5 | 4.8 | 7.8 | **8.7** | **9.7** | 33.3 | **10.8** | 16.2 | 16.8 | 27.6 | 13.84±2.00 |
| SAFixMatch (Li et al., 2025) | **5.0** | 13.9 | 5.0 | **9.4** | 4.4 | **7.3** | 9.9 | 11.7 | **23.4** | 11.4 | **12.3** | **15.4** | **6.3** | **10.41±0.33** |
| GAMT (Qi et al., 2024) | 20.8 | 49.7 | 113.8 | 41.0 | 36.5 | 25.5 | 79.8 | 103.2 | 135.7 | 181.2 | 77.8 | 40.1 | 106.0 | 77.79±5.98 |
| **TriggerMatch (Ours)** | 7.0 | **8.5** | **4.8** | 11.8 | 4.2 | 8.9 | **6.4** | **6.1** | 35.3 | **9.8** | 14.7 | **15.4** | 6.8 | **10.74±1.37** |
| **10 Labeled Cases** | | | | | | | | | | | | | | |
| SupOnly | 9.7 | 13.6 | 7.5 | 11.1 | 3.6 | 8.5 | 9.3 | 9.7 | 19.5 | 9.7 | 16.6 | 15.3 | 18.2 | 11.71±1.15 |
| FixMatch (Sohn et al., 2020) | **5.6** | 12.4 | 7.3 | 9.4 | **4.0** | 8.2 | 4.7 | 6.7 | 32.6 | 6.9 | **11.5** | 13.3 | 21.1 | 11.06±0.54 |
| FreeMatch (Wang et al., 2023) | 49.6 | 41.4 | 17.1 | 27.4 | 8.0 | 9.2 | 5.5 | 7.9 | 32.3 | 9.0 | 68.4 | 21.4 | 56.2 | 27.18±7.08 |
| SoftMatch (Chen et al., 2023) | 8.5 | 18.8 | 9.6 | 9.7 | 5.2 | 7.7 | 5.2 | 7.5 | 33.0 | 7.7 | 30.0 | 13.8 | 29.1 | 14.29±0.71 |
| CPS (Chen et al., 2021) | 7.2 | **9.3** | 6.4 | 8.8 | **3.1** | **7.5** | 4.9 | 7.0 | **24.6** | 7.7 | 16.0 | **12.8** | 9.4 | **9.59±0.14** |
| ReFixMatch (Nguyen & Yang, 2023) | 6.7 | 11.2 | 7.8 | **8.8** | 10.5 | 7.8 | **4.5** | **6.2** | **22.4** | **6.6** | 14.0 | 13.3 | **8.0** | 9.83±1.38 |
| SAFixMatch (Li et al., 2025) | 7.5 | 19.4 | **6.0** | 9.4 | 5.0 | **7.7** | **4.6** | **6.4** | 28.0 | 7.8 | 16.1 | 13.5 | 14.8 | 11.25±0.10 |
| GAMT (Qi et al., 2024) | 21.9 | 53.1 | 76.5 | 57.9 | 37.4 | 28.7 | 97.7 | 125.4 | 117.9 | 147.3 | 62.3 | 47.3 | 93.3 | 74.37±24.68 |
| **TriggerMatch (Ours)** | **4.7** | **9.5** | **3.3** | **7.9** | 5.6 | 8.6 | 5.8 | 6.6 | 25.0 | **5.6** | **10.7** | **13.2** | **6.1** | **8.65±0.15** |
| **20 Labeled Cases** | | | | | | | | | | | | | | |
| SupOnly | 4.3 | 7.8 | 5.1 | 8.4 | 2.3 | 5.8 | 4.5 | 7.3 | 14.0 | 7.9 | 13.5 | 12.8 | 6.5 | 7.72±0.36 |
| FixMatch (Sohn et al., 2020) | 3.6 | 8.5 | 3.8 | 8.1 | 2.5 | 6.2 | 3.8 | 5.5 | 10.6 | 5.2 | **11.2** | 11.9 | 6.2 | 6.71±0.14 |
| FreeMatch (Wang et al., 2023) | 7.1 | 9.7 | 7.4 | 9.2 | 3.5 | 6.4 | 3.9 | 5.6 | 10.7 | 5.3 | 21.3 | 12.3 | 9.2 | 8.60±1.30 |
| SoftMatch (Chen et al., 2023) | **3.5** | 9.2 | 4.3 | **8.0** | 2.4 | 7.4 | 3.8 | 5.5 | 10.0 | 5.5 | **10.5** | 12.2 | 7.2 | 6.91±0.57 |
| CPS (Chen et al., 2021) | **3.3** | **5.3** | 3.8 | **7.6** | **2.1** | 5.6 | **3.6** | 5.6 | **8.8** | 5.4 | 15.1 | **11.7** | 7.3 | **6.55±0.11** |
| ReFixMatch (Nguyen & Yang, 2023) | 3.8 | **6.2** | **3.3** | 8.6 | 2.4 | 6.2 | 3.8 | **5.0** | 12.5 | 5.6 | 11.3 | 12.0 | **6.1** | **6.68±0.24** |
| SAFixMatch (Li et al., 2025) | 6.2 | 9.4 | **3.6** | 8.0 | 2.6 | **5.5** | **3.5** | 5.4 | **8.2** | **5.0** | 14.5 | 12.0 | 6.1 | 6.93±1.06 |
| GAMT (Qi et al., 2024) | 5.6 | 10.4 | 8.5 | 8.3 | 4.3 | 7.8 | 7.7 | 10.1 | 36.3 | 5.7 | 20.5 | 12.0 | 11.4 | 11.43±0.51 |
| **TriggerMatch (Ours)** | 5.7 | 6.2 | 3.7 | 9.0 | **2.0** | **5.5** | 4.0 | **5.1** | 12.9 | **4.9** | 12.8 | **10.9** | **4.6** | 6.71±0.40 |

Table 12: Per-organ Dice (%)↑ on Synapse with 1, 2, 4 labeled volumes. Per-organ values are averaged over seeds (1024, 2026, 2027); Avg. Dice is mean±std over the same seeds. **Red bold**: best; **blue bold**: second best (among semi-supervised methods).

| Method | Spl | R-Kid | L-Kid | Gall | Eso | Liv | Sto | Aor | IVC | PSV | Pan | RAG | LAG | Avg. Dice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1 Labeled volume** | | | | | | | | | | | | | | |
| SupOnly | 53.5 | 67.9 | 47.1 | 3.4 | 25.3 | 86.9 | 19.1 | 65.3 | 46.8 | 17.2 | 12.6 | 7.5 | 2.5 | 35.00±0.78 |
| FixMatch (Sohn et al., 2020) | **68.6** | **83.0** | 67.4 | 3.3 | 21.6 | 86.7 | 23.0 | **82.8** | 64.3 | 20.6 | 7.0 | 6.7 | 2.8 | 41.37±1.06 |
| FreeMatch (Wang et al., 2023) | 18.8 | 48.2 | 24.1 | 2.1 | 18.4 | 60.1 | 11.7 | 52.2 | 17.6 | 12.3 | 13.9 | **8.0** | 1.9 | 22.24±3.35 |
| SoftMatch (Chen et al., 2023) | 40.8 | 68.9 | 56.4 | **3.5** | 23.7 | 84.5 | 21.8 | 75.9 | 54.2 | 18.0 | 29.9 | 8.0 | 2.6 | 37.55±1.42 |
| CPS (Chen et al., 2021) | 57.0 | **83.2** | **70.4** | 3.3 | **28.8** | **88.9** | 24.1 | **82.0** | **66.3** | 10.5 | 4.3 | 6.0 | 2.7 | 40.57±0.48 |
| ReFixMatch (Nguyen & Yang, 2023) | 62.0 | 78.4 | 65.8 | 3.3 | 20.5 | 85.8 | **27.0** | 79.7 | 56.7 | **24.0** | **31.6** | 5.7 | 2.7 | **41.79±1.43** |
| SAFixMatch (Li et al., 2025) | 57.0 | 77.9 | 69.0 | 3.3 | 28.4 | 85.3 | 24.1 | 80.1 | 62.2 | 10.9 | 6.8 | 5.5 | 2.6 | 39.48±0.68 |
| GAMT (Qi et al., 2024) | 24.1 | 41.1 | 29.9 | 2.6 | 5.1 | 64.0 | 9.8 | 20.3 | 12.0 | 7.2 | 8.3 | 3.3 | **2.9** | 17.73±0.77 |
| **TriggerMatch (Ours)** | **74.4** | 81.9 | **76.3** | **3.7** | **36.8** | **88.2** | **25.4** | 80.6 | **66.5** | **39.0** | **36.7** | **11.3** | **3.0** | **47.99±2.67** |
| **2 Labeled volumes** | | | | | | | | | | | | | | |
| SupOnly | 77.2 | 81.2 | 65.6 | 29.7 | 32.5 | 90.7 | 35.7 | 76.0 | 65.3 | 40.4 | 32.1 | 41.1 | 11.8 | 52.26±0.67 |
| FixMatch (Sohn et al., 2020) | 83.3 | **85.1** | 78.8 | 34.8 | 35.0 | 93.9 | 39.7 | **83.6** | **73.9** | 49.0 | 25.3 | 38.5 | 8.0 | 56.08±1.23 |
| FreeMatch (Wang et al., 2023) | 60.9 | 72.9 | 68.6 | 31.1 | 39.6 | 85.8 | 36.3 | 74.6 | 67.7 | 29.5 | 34.7 | 30.3 | 11.0 | 49.45±6.46 |
| SoftMatch (Chen et al., 2023) | 74.5 | 80.3 | 65.6 | 33.8 | 36.5 | 91.6 | **44.9** | 82.4 | 71.1 | 44.0 | 38.3 | 40.4 | 11.5 | 55.00±1.92 |
| CPS (Chen et al., 2021) | **85.6** | **86.4** | **84.2** | 39.4 | **40.0** | **94.2** | 37.6 | **83.4** | 73.3 | **52.1** | 18.3 | **44.9** | 10.6 | **57.71±0.25** |
| ReFixMatch (Nguyen & Yang, 2023) | 84.2 | 84.8 | 81.5 | **40.5** | 39.8 | 93.4 | **44.3** | 82.0 | **74.8** | 41.1 | 33.9 | 38.9 | 6.8 | 57.39±0.07 |
| SAFixMatch (Li et al., 2025) | 84.5 | 84.6 | 81.5 | 33.4 | 37.9 | **94.0** | 37.5 | 82.8 | 72.4 | **49.2** | 27.7 | 39.0 | 11.2 | 56.61±0.83 |
| GAMT (Qi et al., 2024) | 67.1 | 73.3 | 66.8 | 15.9 | 31.6 | 88.6 | 32.1 | 66.1 | 53.9 | 27.6 | **39.4** | 26.8 | **17.2** | 46.65±7.42 |
| **TriggerMatch (Ours)** | **86.9** | 84.5 | **82.9** | **43.5** | **47.4** | 92.2 | 41.9 | 83.2 | 73.8 | 40.4 | **43.4** | **57.4** | **29.3** | **62.07±0.16** |
| **4 Labeled volumes** | | | | | | | | | | | | | | |
| SupOnly | 83.2 | 85.6 | 84.5 | 52.7 | 53.9 | 92.5 | 60.4 | 82.3 | 75.1 | 57.5 | 52.4 | 48.4 | 35.6 | 66.48±0.51 |
| FixMatch (Sohn et al., 2020) | 87.1 | 86.0 | 87.0 | 53.3 | 53.5 | **94.7** | 67.5 | 84.4 | 77.2 | 59.0 | 55.9 | 52.5 | 40.7 | 69.14±0.31 |
| FreeMatch (Wang et al., 2023) | 81.7 | 83.7 | 78.9 | 51.2 | **59.3** | 91.6 | 61.9 | 77.7 | 71.4 | 46.9 | 52.0 | 50.6 | **46.6** | 65.66±1.50 |
| SoftMatch (Chen et al., 2023) | 82.8 | 86.5 | 86.5 | 52.0 | 56.9 | 93.9 | 63.1 | 83.4 | 75.8 | 59.2 | 56.8 | 49.5 | 41.6 | 68.30±1.37 |
| CPS (Chen et al., 2021) | **87.6** | **87.8** | **87.6** | 54.8 | 54.7 | **94.9** | **67.8** | **84.8** | **78.4** | **60.9** | **59.2** | **55.3** | 42.4 | **70.48±0.55** |
| ReFixMatch (Nguyen & Yang, 2023) | **87.4** | 87.0 | 84.6 | **54.9** | 58.4 | 94.2 | 66.6 | 84.3 | 76.8 | 58.5 | 58.3 | 53.2 | 42.1 | 69.71±1.89 |
| SAFixMatch (Li et al., 2025) | 87.0 | **87.1** | 86.6 | 53.9 | 57.6 | 94.5 | **69.0** | 84.3 | 77.8 | **60.8** | 58.3 | 48.3 | 39.4 | 69.58±0.33 |
| GAMT (Qi et al., 2024) | 82.2 | 82.4 | 73.9 | 45.7 | 56.7 | 93.4 | 63.5 | 78.5 | 73.3 | 49.6 | 57.2 | 53.3 | 45.8 | 65.82±3.19 |
| **TriggerMatch (Ours)** | 85.6 | 85.7 | **87.3** | **56.1** | **62.9** | 94.5 | 67.5 | **84.6** | **78.4** | 58.6 | **64.7** | **61.3** | **61.0** | **72.94±0.93** |

Table 13: Per-organ HD95 (mm)↓ on Synapse with 1, 2, 4 labeled volumes. Per-organ values are averaged over seeds (1024, 2026, 2027); Avg. HD95 is mean±std over the same seeds. **Red bold**: best; **blue bold**: second best (among semi-supervised methods).

| Method | Spl | R-Kid | L-Kid | Gall | Eso | Liv | Sto | Aor | IVC | PSV | Pan | RAG | LAG | Avg. HD95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1 Labeled volume** | | | | | | | | | | | | | | |
| SupOnly | 130.7 | 59.9 | 49.3 | 88.2 | 23.3 | 63.5 | 72.4 | 33.2 | 50.8 | 91.1 | 78.5 | 42.8 | 81.3 | 66.53±6.01 |
| FixMatch (Sohn et al., 2020) | **89.7** | **15.4** | **25.4** | 86.0 | 28.6 | 71.1 | 68.5 | **17.9** | **20.0** | **61.4** | 71.7 | 35.9 | 50.6 | **49.41**±5.23 |
| FreeMatch (Wang et al., 2023) | 164.2 | 132.2 | 125.7 | 134.4 | 61.0 | 125.8 | 127.9 | 81.3 | 122.4 | 112.3 | 91.9 | 68.5 | 103.9 | 111.65±16.30 |
| SoftMatch (Chen et al., 2023) | 161.9 | 61.1 | 58.0 | 87.1 | 29.3 | 84.7 | 102.4 | **24.4** | 55.0 | 83.3 | 63.1 | 50.9 | 61.8 | 70.99±2.88 |
| CPS (Chen et al., 2021) | 97.0 | **11.6** | **16.2** | 81.1 | 21.9 | **51.7** | **53.6** | 25.1 | **16.8** | 83.0 | 94.5 | **34.3** | 43.3 | **48.46**±2.14 |
| ReFixMatch (Nguyen & Yang, 2023) | 100.7 | 43.9 | 84.3 | **77.0** | 24.8 | 62.1 | 76.7 | 38.7 | 64.4 | **50.5** | **53.7** | 42.6 | **37.4** | 58.21±11.03 |
| SAFixMatch (Li et al., 2025) | 129.3 | 50.6 | 34.5 | 90.9 | **19.9** | 72.6 | **54.2** | 42.0 | 21.9 | 67.2 | 84.1 | **29.2** | 80.7 | 59.78±2.03 |
| GAMT (Qi et al., 2024) | 182.3 | 149.8 | 145.8 | 143.9 | 149.4 | 131.7 | 127.4 | 136.7 | 161.2 | 133.0 | 124.7 | 148.6 | 143.5 | 144.46±4.27 |
| **TriggerMatch (Ours)** | **88.3** | 30.4 | 37.7 | **55.2** | **17.7** | **60.7** | 60.2 | 49.2 | 54.8 | 65.2 | **52.9** | 57.8 | **32.2** | 50.94±7.74 |
| **2 Labeled volumes** | | | | | | | | | | | | | | |
| SupOnly | 55.4 | 9.7 | 25.0 | 44.7 | 21.9 | 16.0 | 38.2 | 18.3 | 12.4 | 38.3 | 34.0 | 9.7 | 22.5 | 26.63±4.25 |
| FixMatch (Sohn et al., 2020) | 34.0 | 13.1 | **13.1** | 37.5 | 16.5 | 11.6 | 37.3 | **16.3** | 11.8 | 36.9 | 40.0 | **9.0** | **23.2** | 23.11±1.08 |
| FreeMatch (Wang et al., 2023) | 103.0 | 81.8 | 50.5 | 81.3 | 30.4 | 67.8 | 73.4 | 42.9 | 33.6 | 53.6 | 43.7 | 23.5 | 43.4 | 56.07±21.08 |
| SoftMatch (Chen et al., 2023) | 76.4 | 17.1 | 19.1 | 41.6 | 25.9 | 14.0 | 55.0 | 18.8 | 11.5 | 41.6 | 36.1 | 11.8 | 27.4 | 30.49±3.70 |
| CPS (Chen et al., 2021) | **25.3** | **6.7** | **5.4** | **25.5** | **14.7** | **8.1** | **34.9** | 17.3 | **8.2** | 43.0 | 48.4 | **8.4** | 24.1 | **20.77**±0.91 |
| ReFixMatch (Nguyen & Yang, 2023) | 37.0 | 18.2 | 20.6 | 44.6 | 17.2 | 13.7 | 46.1 | 21.5 | 13.6 | 40.3 | **32.7** | 11.3 | 33.0 | 26.90±1.09 |
| SAFixMatch (Li et al., 2025) | 27.7 | **8.5** | 16.2 | **30.2** | 17.0 | **7.5** | 36.8 | **16.6** | **7.1** | **36.0** | 37.7 | 10.0 | 27.5 | 21.44±0.57 |
| GAMT (Qi et al., 2024) | 121.7 | 96.6 | 82.7 | 135.8 | 88.0 | 52.8 | 94.8 | 72.4 | 85.3 | 69.4 | 60.6 | 84.2 | 91.6 | 87.38±23.10 |
| **TriggerMatch (Ours)** | **22.5** | 11.4 | 15.9 | 30.5 | **13.9** | 14.0 | **31.8** | 22.1 | 11.2 | **28.6** | **22.5** | 10.4 | **14.7** | **19.20**±2.63 |
| **4 Labeled volumes** | | | | | | | | | | | | | | |
| SupOnly | 49.5 | 13.2 | 17.6 | 42.9 | 10.4 | 10.3 | 25.8 | 17.2 | 8.8 | 18.7 | 20.3 | 11.4 | 12.5 | 19.90±1.77 |
| FixMatch (Sohn et al., 2020) | 25.1 | 15.0 | 11.4 | 24.6 | 13.4 | 9.9 | **22.9** | 15.1 | **5.8** | 15.3 | 12.6 | **5.9** | 9.5 | 14.35±1.75 |
| FreeMatch (Wang et al., 2023) | 70.7 | 38.5 | 40.9 | 46.8 | 11.1 | 29.1 | 60.4 | 28.5 | 34.9 | 33.4 | 36.6 | 12.9 | 11.2 | 35.00±0.89 |
| SoftMatch (Chen et al., 2023) | 53.4 | 18.7 | 11.5 | 22.4 | 12.2 | 13.3 | 34.2 | 14.8 | 7.2 | 19.0 | 18.1 | 8.0 | 8.6 | 18.57±0.65 |
| CPS (Chen et al., 2021) | **20.3** | **4.7** | **7.9** | 25.2 | 12.9 | **6.0** | **19.9** | 16.5 | 6.3 | **14.7** | 15.4 | 7.8 | **8.6** | **12.78**±1.45 |
| ReFixMatch (Nguyen & Yang, 2023) | 31.6 | 8.4 | 14.3 | **21.2** | **10.8** | 9.7 | 31.4 | **14.6** | 9.6 | 15.1 | 16.7 | 7.7 | 8.9 | 15.38±2.63 |
| SAFixMatch (Li et al., 2025) | 32.3 | **7.9** | 12.9 | **17.7** | 11.9 | 9.1 | 23.1 | **14.4** | **5.8** | 15.1 | **12.3** | 8.8 | 9.2 | 13.88±1.44 |
| GAMT (Qi et al., 2024) | 63.9 | 26.7 | 55.5 | 62.8 | 22.2 | 21.3 | 41.6 | 20.5 | 16.1 | 31.3 | 28.6 | 9.2 | 22.3 | 32.47±15.45 |
| **TriggerMatch (Ours)** | **24.3** | 10.5 | **7.1** | 25.0 | **8.4** | **5.5** | 24.4 | 21.9 | 11.6 | **14.8** | **11.3** | **5.2** | **6.7** | **13.59**±2.27 |

Table 14: Controller hyperparameters. The same values are used in every dataset and label regime.

| Layer | Symbol | Role | Value |
|---|---|---|---|
| Measurement | $\tau_a$ | audit confidence | 0.95 |
| | $\beta$ | recall-EMA decay | 0.96 |
| Decision | $\rho_{\text{in}}$ / $\rho_{\text{out}}$ | entrance / exit recall | 0.95 / 0.85 |
| | $K_{\text{in}}$ / $K_{\text{out}}$ / $K_{\text{np}}$ | patience | 5 / 5 / 5 |
| | $T_w$ | warm-up | 5 epochs |
| Branch | $\lambda_u$ | unlabeled weight | 1.0 |

Table 15: Optimization settings, shared verbatim by TriggerMatch and all baselines. $T$ denotes the total iteration count of Table 7.

| Component | Setting | Value |
|---|---|---|
| Optimizer | algorithm | AdamW |
| | initial learning rate | 0.1 |
| | betas $(\beta_1, \beta_2)$ | $(0.9,\ 0.999)$ |
| | weight decay | $10^{-3}$ |
| | $\epsilon$ | $10^{-8}$ (default) |
| Schedule | decay rule | polynomial, per iteration $\text{lr}_t = \text{lr}_0\,(1 - t/T)^{0.9}$ |
| Batching | crop size | 64×128×128 |
| | labeled / unlabeled crops | $B_l$ / $B_l$ |
| | batch size | 8 |

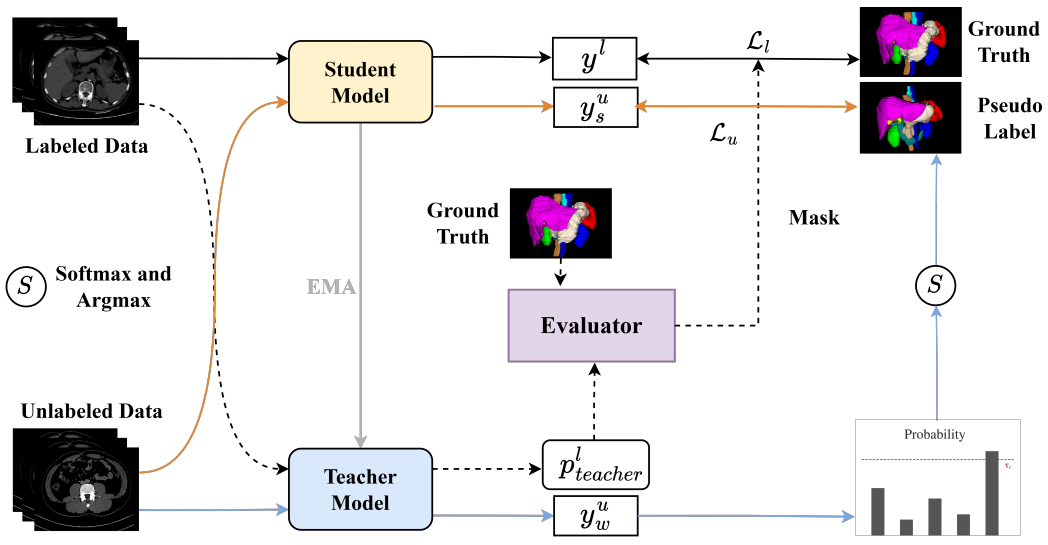



Figure 7: Overall framework of TriggerMatch. Solid black lines denote the training path of the labeled data; orange lines, the training path of strongly augmented unlabeled data; blue lines, the training path of weakly augmented unlabeled data; dashed lines denote the verification path along which the semantic understanding of the teacher that provides the pseudo-labels is audited.