SPEGC: Continual Test-Time Adaptation via Semantic-Prompt-Enhanced Graph Clustering for Medical Image Segmentation

Supplementary Material

A. Theoretical Analysis

In the main paper(Sec. 3.3), we state that the refined edge similarity matrix, S^*, "approximates global consistency." This appendix provides the formal mathematical proof for this claim. We demonstrate that S^* is, by necessity, a differentiable approximation rather than a strict satisfaction of global consistency. This approximation is a direct and necessary consequence of reframing a discrete combinatorial problem into a continuous, end-to-end differentiable optimization landscape.

The objective of our Differentiable Graph Clustering Solver (DGCS) is to extract a refined, continuous affinity structure. While motivated by the fact that a discrete spanning forest with Z components contains $k = V - Z$ edges, DGCS relaxes this discrete partitioning. Instead, it utilizes k as a sparsity budget to perform a differentiable global k-edge sparsification via an Optimal Transport (OT) relaxation.

Definition 1 (Discrete Global Consistency)**.** *We define "true" or "hard" global consistency as the discrete, combinatorial solution to this graph partitioning problem. This solution is a binary assignment vector $y \in \{0,1\}^E$ (where $E = V^2$ is the total number of possible edges) such that $\sum_{i=1}^{E} y_i = k$, and y maximizes the total affinity defined by S'. Note that while this definition selects k edges to maximize affinity, our OT relaxation does not explicitly enforce the acyclicity required for a true spanning forest, yielding instead a probabilistic soft affinity matrix S^*.*

This discrete selection process, $y = \text{argmax}(\dots)$, is non-differentiable and thus cannot be used for gradient-based optimization in an end-to-end framework.

To bridge this gap, we reformulate the problem as an optimal transport (OT) task, as introduced in Eqs. (11) to (13). The core objective is to find a transport plan $\Gamma \in \mathbb{R}^{E\times 2}$ that maps E edges to two "bins": "reject" or "select", while matching the marginal constraints $r = \mathbf{1}_E$ and $c = [E-k, k]^T$.

Let us first consider the standard, *unregularized* OT problem (i.e., a classic Linear Program):

$$\Gamma^*_{\text{hard}} = \arg\min_{\Gamma} \langle \Gamma, D \rangle \quad \text{s.t.} \quad \Gamma \mathbf{1}_2 = r, \Gamma^T \mathbf{1}_E = c$$

where D is the cost matrix defined in Eqs. (11) and (12).

Lemma 1 (Nature of the Unregularized Solution)**.** *The solution Γ^*_{hard} to this unregularized linear program is guaranteed to be a "hard" assignment. According to the Birkhoff-von Neumann theorem, the vertices of the feasible polytope of Γ (the set of doubly-stochastic matrices, or in our case, matrices with fixed marginals) correspond to permutation matrices (or binary assignments). Therefore, the solution Γ^*_{hard} would represent a discrete, binary selection of exactly k edges. The second column, $\Gamma^*_{hard,:,2}$, would be the exact binary vector y from Definition 1.*

While Γ^*_{hard} would satisfy "true" global consistency, solving this linear program is (a) computationally expensive and (b) the solution process is non-differentiable with respect to the input cost matrix D.

The critical step in our DGCS is the introduction of the **entropy regularization** term, $\theta h(\Gamma)$, as shown in Eq. (13):

$$\Gamma^* = \arg\min_{\Gamma} \langle \Gamma, D \rangle + \theta h(\Gamma)$$

where $h(\Gamma) = -\sum_{i,j} \Gamma_{i,j}(\log(\Gamma_{i,j}) - 1)$ is the entropy, and θ is the regularization temperature.

Theorem 1 (Consequence of Entropy Regularization)**.** *The introduction of the strictly convex entropy term $h(\Gamma)$ (for $\theta > 0$) achieves two goals:*

1. *It makes the objective function strictly convex, guaranteeing a* unique *solution Γ^*.*
2. *It enables the use of the highly efficient, parallel, and—most importantly—**differentiable** Sinkhorn-Knopp algorithm for finding the solution Γ^*.*

However, this unique solution Γ^* is no longer the discrete Γ^*_{hard}. The entropy term "softens" the assignment, penalizing sparse (binary) solutions and favoring "diffuse" solutions. The resulting Γ^* is a "soft" matrix where entries $\Gamma^*_{i,j} \in (0,1)$, not $\{0,1\}$.

This Γ^* is a differentiable *approximation* of the "hard" linear programming solution Γ^*_{hard}. The regularization parameter θ explicitly controls this trade-off:

- As $\theta \to 0$, the problem converges to the "hard" (non-differentiable) linear program, and $\Gamma^* \to \Gamma^*_{\text{hard}}$.
- As $\theta \to \infty$, the entropy term dominates, and the solution ignores the cost D, becoming $\Gamma^*_{i,j} \propto r_i c_j$.

By using a finite, non-zero θ, we explicitly choose to operate in the "soft," approximate regime in order to gain differentiability.

In SPEGC, the final refined edge similarity matrix S^* is constructed by reshaping the second column of this *soft* transport plan: $S^* = \text{reshape}(\Gamma^*_{:,2})$.

Since Γ^* is a "soft" matrix of non-binary values, its second column $\Gamma^*_{:,2}$ is not a binary vector of k ones and $E-k$

Algorithm 1 SPEGC: Continual Test-Time Adaptation via Semantic-Prompt-Enhanced Graph Clustering

Initialize: Source-trained model f_σ with parameters $\sigma \leftarrow \sigma_S$;
Learnable prompt pools $P_{CO} \in \mathbb{R}^{M\times h}$, $P_{HE} \in \mathbb{R}^{M\times h}$;
Learnable projections $W_q \in \mathbb{R}^{h\times h}, W_k \in \mathbb{R}^{h\times h}$;
Learnable context vector $c_p \in \mathbb{R}^h$;
Feature queue $\mathcal{Q} \leftarrow \emptyset$ (max capacity N);
Input: Continuous target domain image stream $\{x_i\}_{i=1}^{\infty}$.
Output: Adapted prediction stream $\{O_i\}_{i=1}^{\infty}$.
for $i \in [1, m]$ **do**
1. Semantic Prompt Feature Enhancement (SPFE)
  Generate t stochastic feature maps $\{F_k\}_{k=1}^{t} \leftarrow f_\sigma(x_i)$ using MC Dropout.
  Estimate uncertainty $U \leftarrow \text{Variance}(\{F_k\}_{k=1}^{t})$ (Eq. (1)).
  Select n_i low-uncertainty nodes $\mathcal{V}_i \leftarrow \text{SampleLowUncertainty}(U, \{F_k\}, p\%)$.
  Aggregate node features into global query $\hat{q}_i \leftarrow \text{AttentionPooling}(\mathcal{V}_i, c_p)$ (Eq. (2)).
  Retrieve commonality prompt $p_{CO}(i) \leftarrow \text{ReverseAttention}(\hat{q}_i, P_{CO})$ (Eqs. (3) and (5)).
  Retrieve heterogeneity prompt $p_{HE}(i) \leftarrow \text{Attention}(\hat{q}_i, P_{HE})$ (Eqs. (4) and (6)).
  Obtain enhanced features $\mathcal{V}_i^* \leftarrow \mathcal{V}_i + p_{CO}(i) + p_{HE}(i)$ (Eq. (7)).
2. Differentiable Graph Clustering Solver (DGCS)
  Assemble pseudo-batch $\mathcal{V}^* \leftarrow \mathcal{Q} \cup \{\mathcal{V}_i^*\}$.
  Update feature queue: $\mathcal{Q}$.enqueue$(\mathcal{V}_i^*)$.
  if $|\mathcal{Q}| > N$ **then**
    $\mathcal{Q}$.dequeue$()$.
  Calculate global similarity $S \in \mathbb{R}^{V\times V}$ (Eq. (8)), where V is total nodes in $\mathcal{V}^*$.
  Determine node densities $D(v_i) \leftarrow \sum_j \text{ReLU}(S(i,j))$ (Eq. (9)).
  Define density-aware edge similarity $S'(i,j)$ (Eq. (10)).
  Formulate optimal transport cost matrix $\mathbf{D}$ from S' (costs for selecting/rejecting $k = V - Z$ edges) (Eqs. (11) and (12)).
  Solve for optimal transport plan $\Gamma^* \leftarrow \text{Sinkhorn}(\mathbf{D}, \theta)$ (Eqs. (13) and (14)).
  Extract refined edge similarity $S^* \leftarrow \text{Reshape}(\Gamma^*_{:,2})$.
3. Joint Optimization & Adaptation
  Acquire semantic predictions $P \in \mathbb{R}^{V\times C}$ for nodes in $\mathcal{V}^*$ from f_σ.
  Calculate graph consistency loss L_G (Eq. (16)) and clustering loss L_C (Eq. (17)).
  Determine total loss $L \leftarrow L_G + \lambda L_C$ (Eq. (15)).
  Update all learnable parameters $\{\sigma, P_{CO}, P_{HE}, W_q, W_k, c_p\}$ by backpropagating L.
4. Inference
  Generate final prediction $O_i \leftarrow f_\sigma(x_i)$ (using the updated parameters σ).
end for

Table A.1. Quantitative comparison under **Mixed Distribution Shifts** on the OD/OC segmentation task. Models are trained on the indicated source domain and adapted to a composite target stream formed by shuffling samples from all remaining domains. We report the Mean ± Std. over five independent runs. The best results are highlighted in **bold red**.

Methods	Domain A			Domain B			Domain C			Domain D			Domain E			Average		
	DSC	$E_\phi^{\max}$	S_α	DSC	$E_\phi^{\max}$	S_α	DSC	$E_\phi^{\max}$	S_α	DSC	$E_\phi^{\max}$	S_α	DSC	$E_\phi^{\max}$	S_α	DSC ↑	$E_{\max}^{b}$ ↑	S_α ↑
No Adapt (ResUNet-50) [10]	71.92	88.67	80.84	79.31	90.42	84.82	75.42	89.24	81.62	63.77	85.49	78.28	73.32	89.67	81.81	72.75	88.70	81.47
SAR(ICLR'23) [46]	74.03±6.43	90.07±0.31	83.52±0.17	80.33±2.42	92.86±0.31	85.29±0.19	71.42±4.67	91.55±0.12	83.01±0.16	69.89±1.32	86.01±0.14	79.17±0.98	69.75±5.61	88.03±0.97	86.12±0.47	73.08	89.70	83.42
Domain Adaptor(CVPR'23) [69]	76.28±4.47	90.75±0.40	83.82±0.14	76.27±4.28	91.03±0.37	85.71±0.14	70.21±8.12	91.01±0.31	82.44±0.24	66.18±9.82	83.41±0.42	78.34±0.17	76.31±4.42	88.71±0.21	84.62±0.08	73.05	88.98	82.99
NC-TTT(CVPR'24) [48]	76.81±2.12	92.79±0.23	85.69±0.13	82.74±3.48	**93.47±0.34**	**86.71±0.10**	77.93±6.07	92.26±0.21	84.02±0.14	75.05±4.12	87.74±0.31	81.54±0.39	81.23±0.39	92.61±0.14	85.01±0.12	78.75	91.77	84.59
VPTTA(CVPR'24) [5]	75.17±5.01	92.44±0.10	85.71±0.01	79.02±3.94	92.01±0.07	84.84±0.06	71.41±2.56	92.24±0.10	82.79±0.06	64.02±6.02	85.89±0.14	77.61±0.11	75.73±3.41	90.72±0.14	84.41±0.11	73.07	90.66	83.08
GraTA(AAAI'25) [6]	78.14±4.49	91.98±0.08	83.51±0.10	81.21±3.03	91.71±0.27	84.14±0.14	77.02±3.43	91.57±0.41	83.89±0.42	74.15±3.78	90.15±0.34	82.46±0.17	74.79±3.57	88.62±0.18	82.00±0.13	77.06	90.08	83.20
TTDG(CVPR'25) [39]	82.74±3.10	**94.02±0.14**	86.81±0.19	**82.91±3.42**	92.09±0.21	85.99±0.12	82.97±2.14	**93.47±0.32**	87.00±0.18	78.74±4.07	91.14±0.28	83.97±0.10	84.51±3.27	**93.64±0.27**	87.12±0.14	82.37	92.82	86.18
SPEGC(Ours)	**84.72±2.30**	93.64±0.47	**88.09±0.24**	82.88±1.79	92.89±0.26	86.41±0.37	**83.79±2.41**	93.42±0.40	**88.02±0.17**	**83.04±2.57**	**93.04±0.11**	**85.32±0.14**	**85.02±2.07**	93.34±0.28	**88.81±0.13**	**83.89**	**93.47**	**87.33**

zeros. Instead, it is a vector of **soft probabilities** or "selection likelihoods" for each edge.

Therefore, S^* is not a "hard" adjacency matrix representing a discrete, globally consistent spanning forest. It is, by its very mathematical construction, a **differentiable approximation** of that ideal structure.

The "approximation" of global consistency is the necessary trade-off for the "differentiability" of the clustering solver. This soft structural representation enables gradient flow via graph consistency loss for end-to-end adaptation.

Table A.2. Quantitative comparison under **Mixed Distribution Shifts** on the polyp segmentation task. Models are trained on the indicated source domain and adapted to a composite target stream formed by shuffling samples from all remaining domains. We report the Mean ± Std. over five independent runs. The best results are highlighted in **bold red**.

Methods	Domain A			Domain B			Domain C			Domain D			Average		
	DSC	E^b_{max}	S_α	DSC	E^b_{max}	S_α	DSC	E^b_{max}	S_α	DSC	E^b_{max}	S_α	DSC ↑	E^b_{max} ↑	S_α ↑
No Adapt (ResUNet-50) [10]	70.32	86.82	82.14	68.33	85.33	81.32	70.48	87.38	81.92	76.81	87.19	84.14	71.49	86.68	82.38
SAR(ICLR'23) [46]	69.01±0.98	85.79±0.14	80.21±0.14	67.81±2.32	83.87±0.07	81.51±0.14	70.51±1.47	88.31±0.17	82.40±0.09	68.02±2.44	84.37±0.14	79.68±0.06	68.84	85.59	80.95
Domain Adaptor(CVPR'23) [69]	78.17±1.14	91.84±0.07	86.41±0.12	70.87±1.74	89.36±0.05	82.13±0.12	71.63±2.04	91.98±0.12	83.04±0.04	68.76±1.14	83.97±0.07	79.52±0.04	72.36	89.29	82.78
NC-TTT(CVPR'24) [48]	77.64±1.31	91.54±0.17	86.52±0.12	**73.51±2.34**	**91.64±0.04**	**83.17±0.04**	68.93±1.02	89.81±0.17	81.63±0.12	80.67±1.04	89.51±0.14	84.28±0.13	75.19	90.63	83.90
VPTTA(CVPR'24) [5]	77.38±0.68	92.11±0.09	86.76±0.10	71.48±0.87	88.42±0.07	81.97±0.13	71.01±0.82	91.64±0.08	81.93±0.10	80.42±0.29	89.93±0.11	84.54±0.07	75.07	90.53	83.80
GraTA(AAAI'25) [6]	78.67±3.78	92.54±0.19	87.81±0.21	68.82±2.62	87.97±0.19	81.54±0.12	**72.81±3.77**	92.22±0.21	82.82±0.21	81.52±2.96	89.33±0.21	85.19±0.14	75.45	90.51	84.54
TTDG(CVPR'25) [39]	82.21±1.46	93.69±0.11	**89.77±0.19**	70.02±1.64	88.72±0.12	81.13±0.06	69.82±2.37	91.63±0.14	83.08±0.08	80.54±1.27	89.14±0.05	85.16±0.11	75.65	90.78	84.74
SPEGC(Ours)	**83.04±1.14**	**94.03±0.08**	89.72±0.09	72.62±0.86	89.06±0.14	82.17±0.08	72.51±0.63	**92.61±0.08**	**84.32±0.04**	**83.82±0.93**	**90.48±0.11**	**86.32±0.09**	**78.00**	**91.55**	**85.68**

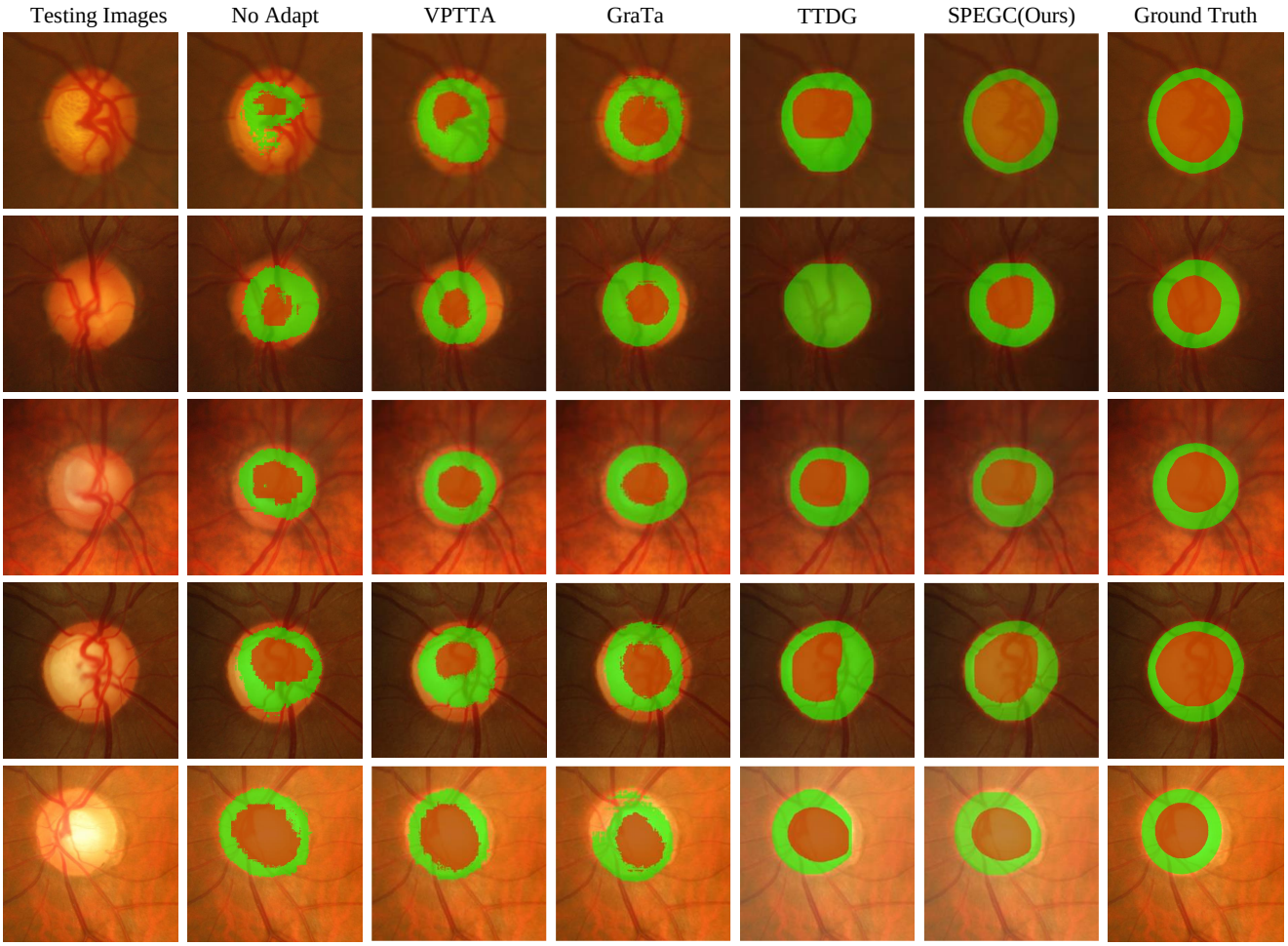

Figure A.1. Visualization comparison of segmentation results for the No Adapt baseline, VPTTA [5], GraTa [6], TTDG [39], and SPEGC(Ours) in retinal fundus segmentation. Different colors represent the segmentation instances of different classes identified by the network.

B. Algorithm Pipeline

Algorithm 1 outlines the comprehensive workflow of SPEGC framework. It details the per-sample continual adaptation procedure, which is structured into three principal stages: (1) Semantic Prompt Feature Enhancement (SPFE), (2) Differentiable Graph Clustering Solver (DGCS), (3) Joint Optimization & Adaptation.

C. More Experiments

C.1. Result Visualization

To further validate SPEGC, we provide qualitative comparisons for the continual test-time adaptation (CTTA) stream on both retinal fundus (OD/OC) and polyp segmentation tasks, presented in Fig. A.1 and Fig. A.2, respectively. Each

Figure A.2. Visualization comparison of segmentation results for the No Adapt baseline, VPTTA [5], GraTa [6], TTDG [39], and SPEGC(Ours) in polyp segmentation.

row visualizes the model's adaptive performance on a distinct, unseen target domain, simulating the challenging real-world scenario of evolving data distributions. The OD/OC segmentation (Fig. A.1) is inherently difficult, demanding precise delineation of two overlapping anatomical structures (Optic Disc and Cup) often obscured by low contrast and domain-specific artifacts. The polyp segmentation (Fig. A.2) poses an even greater challenge due to extreme inter-domain variance in lesion morphology, including drastic differences in shape, size, and texture.

As visualized Tab. 2, many competing CTTA methods exhibit significant performance degradation as the domain stream progresses. They suffer from error accumulation or catastrophic forgetting, resulting in noisy predictions, incomplete structures, or a collapse towards source-domain priors. In stark contrast, SPEGC maintains robust and accurate segmentation. This is primarily attributed to our two-fold mechanism: (1) SPFE injects robust global contextual information, effectively mitigating feature-level noise induced by the domain shift and preserving cross-domain commonalities. (2) DGCS distills a refined, high-order structural representation from these enhanced features. This structure acts as a stable, cluster-level supervisory signal, guiding the model to adapt without compromising core semantic knowledge. As evidenced in the figures, SPEGC consistently produces precise and structurally coherent segmentation masks that closely align with ground-truth annotations, demonstrating superior stability and adaptation fidelity across diverse and evolving domains, particularly in scenarios where other methods exhibit instability.

C.2. Comparison experiments under mixed distribution shifts

To better emulate the complexity of real-world clinical environments, where test data frequently arrive in arbitrarily mixed and continuously evolving streams, we conducted

a rigorous evaluation under Mixed Distribution Shifts. In this protocol, a source model trained on a single domain is adapted to a composite target stream, constructed by shuffling samples from all remaining target domains. For reproducibility, the random seed for all data shuffling procedures was fixed at 2026. Quantitative comparisons on the OD/OC and polyp segmentation benchmarks are reported in Tab. A.1 and Tab. A.2, respectively. As evidenced by the results, SPEGC consistently outperforms SOTA methods, achieving the highest average DSC scores of **83.89%** and **78.00%** across the two tasks. Notably, compared to the runner-up TTDG [39], SPEGC exhibits superior stability across varying domains. This highlights the structural robustness and generalization capability of SPEGC in handling complex, online CTTA scenarios.

C.3. Additional Evaluation Metric: ASSD

While the Dice Similarity Coefficient (DSC) effectively measures the regional overlap, we additionally introduce the Average Symmetric Surface Distance (ASSD) to further rigorously evaluate the boundary delineations of the segmentation predictions. The ASSD results for the existing 2D tasks (retinal fundus and polyp segmentation) are presented in Tab. A.3 and Tab. A.4. For ASSD, we report the metric derived from a single run, where the standard deviation (Std.) is computed across the test images. As observed, SPEGC demonstrates highly competitive performance across virtually all comparisons, indicating that our structural refinement not only improves semantic overlap but also achieves superior boundary consistency.

Table A.3. ASSD (in pixels) for the OD/OC segmentation task. Red and blue indicate the best and second-best results, respectively.

Methods	Domain A	Domain B	Domain C	Domain D	Domain E
No Adapt	46.17	37.14	39.82	54.27	44.72
SAR	42.28±30.71	35.81±26.93	44.59±32.47	48.90±27.19	47.31±30.72
Domain Adaptor	40.54±27.54	38.17±29.92	46.15±31.72	51.23±33.02	36.28±20.39
NC-TTT	38.94±28.19	**24.82±20.41**	36.54±26.35	40.62±28.40	30.29±21.79
VPTTA	45.28±32.54	34.33±26.17	46.21±29.86	53.81±34.92	39.34±30.83
GraTA	34.62±24.47	29.74±23.22	35.77±27.58	38.04±28.49	37.81±26.04
TTDG	**27.61±17.33**	**25.16±19.34**	**29.06±21.34**	**33.97±25.71**	**23.63±18.49**
SPEGC (Ours)	**29.30±19.67**	25.71±16.57	**26.33±18.37**	**27.64±22.10**	**24.08±20.71**

Table A.4. ASSD (in pixels) for the polyp segmentation task. Red and blue indicate the best and second-best results, respectively.

Methods	Domain A	Domain B	Domain C	Domain D
No Adapt	30.49	34.36	32.19	27.02
SAR	27.70±12.71	35.17±16.97	32.04±21.40	34.83±15.60
Domain Adaptor	23.19±11.94	29.76±14.57	32.49±19.92	33.91±17.42
NC-TTT	24.72±14.62	**25.57±14.33**	33.12±24.63	27.43±13.81
VPTTA	23.91±12.02	27.03±19.72	31.99±25.72	29.07±16.03
GraTA	22.54±16.55	30.11±18.06	**30.54±23.99**	26.39±15.44
TTDG	**20.34±13.82**	27.69±13.47	33.72±21.46	27.10±14.52
SPEGC (Ours)	**21.17±12.13**	**26.24±14.09**	**31.27±20.51**	**25.71±13.84**

C.4. Extension to 3D Medical Image Segmentation

To further validate the robustness and applicability of our method across diverse modalities and spatial dimensions, we extend our evaluation to 3D volumetric data using the M&MS (Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation) dataset (MRI modality). Following the adaptation paradigm explored in SicTTA [63], we report both DSC and ASSD for this 3D task.

As shown in Tab. A.5 and Tab. A.6, SPEGC maintains significant performance gains over the baseline on the 3D M&MS dataset. Note that for DSC, we conduct five independent runs and report the standard deviation across these runs. A direct quantitative comparison with SicTTA is omitted in these tables due to differences in the underlying backbone architectures; whereas our baseline uniformly employs ResUNet-50, SicTTA utilizes a different backbone network.

Table A.5. DSC (Mean ± Std.) performance of SPEGC on the 3D M&MS dataset. Red indicates the best results.

Methods	Domain A			Domain B			Domain C			Domain D		
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
NoAdapt	83.72	70.19	72.34	76.29	69.92	68.57	77.91	68.67	65.07	78.39	68.37	64.72
SPEGC	**87.64±1.86**	**79.80±2.72**	**75.28±2.17**	**84.16±1.54**	**78.49±2.63**	**75.34±2.09**	**86.49±1.68**	**76.10±1.82**	**70.62±2.11**	**85.17±1.20**	**74.39±2.17**	**69.32±2.45**

Table A.6. ASSD (pixels) (Mean ± Std.) performance of SPEGC on the 3D M&MS dataset. Red indicates the best results.

Methods	Domain A			Domain B			Domain C			Domain D		
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
NoAdapt	4.37	4.61	4.59	4.82	5.07	5.11	4.39	5.16	5.71	5.38	4.80	4.91
SPEGC	**3.76±3.37**	**4.04±3.92**	**4.21±3.47**	**4.17±3.88**	**4.39±4.01**	**4.82±3.90**	**3.26±3.07**	**4.37±3.52**	**5.02±4.31**	**4.14±2.28**	**4.06±3.16**	**4.72±3.75**

C.5. Sensitivity Analysis of Hyperparameter λ

To further investigate the impact of the loss balancing coefficient λ (defined in Eq. (15) of the main paper) on the adaptation performance, we conduct an additional sensitivity analysis on the OD/OC segmentation task. As presented in Tab. A.7, SPEGC demonstrates robust stability across a range of values, with the average DSC peaking at $\lambda = 0.2$.

Notably, when $\lambda = 0$, SPEGC relies solely on the graph consistency loss L_G, yielding sub-optimal results (76.83%) due to the lack of explicit semantic constraints on the commonality prompt pool. Conversely, performance degrades slightly at higher values (e.g., $\lambda \geq 0.6$) as the clustering loss L_C dominates the optimization, potentially overshadowing the structural guidance provided by L_G. Consequently, we adopt $\lambda = 0.2$ as the optimal default configuration for all experiments.

Table A.7. Ablation study of the hyperparameter λ on the OD/OC segmentation task (Average DSC). Red indicates the best result.

Metric	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.8$
DSC (%)	76.83	81.60	**84.37**	83.79	82.14	80.42	79.59