Hi-FL: Hierarchical Federated Learning Adaptation of Vision-Language Models for Multi-Institutional Surgical Phase Recognition

Julia Alekseenko[1,6][0009−0007−4698−8957], Giuseppe Quero[2], Giovanni Guglielmo Laracca[3], Ludovica Baldari[4], Salvador Morales-Conde[5], Pietro Mascagni[6,2], Didier Mutter[6,7], and Nicolas Padoy[1,6]

[1] University of Strasbourg, Strasbourg, France
alekseenko@unistra.fr
[2] Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy
[3] Azienda Ospedaliero-Universitaria Sant'Andrea, Rome, Italy
[4] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
[5] University Hospital Virgen Macarena, University of Sevilla, Seville, Spain
[6] IHU Strasbourg, Strasbourg, France
[7] University Hospital of Strasbourg, Strasbourg, France

Abstract. Automated surgical phase recognition is challenging across decentralized centers due to domain shifts in equipment and technique, along with strict privacy constraints. While pre-trained Vision–Language Models (VLMs) provide strong semantic priors, existing federated VLM adaptation methods remain ineffective because they treat surgical phases as flat categories, mixing center-specific noise with true phase semantics. We propose Hi-FL, a Federated Learning framework for Hierarchical Adaptation (HA) of VLMs, which splits procedural logic into three levels: (i) *procedure-level* global context over the procedure, (ii) *phase-level* dynamics across stages, and (iii) *action-level* frame-wise cues. To handle heterogeneity across centers, we also introduce a Class-Conditional Targeted Alignment and Repulsion (TAR) loss that aligns local features to global semantics while minimizing cross-institutional interference.
Evaluated on Cholecystectomy, Colorectal, and Gastric-Bypass procedures, Hi-FL consistently outperforms federated and few-shot baselines, improving accuracy by up to 10% in moderate-data regimes (5%–50% videos per center) and in low-shot settings (16–64 frames). Additionally, it preserves semantic phase structure and merges multi-center distributions, overcoming center-specific noise, making it a practical framework for privacy-preserving, cross-site deployment of VLMs in surgery. Implementation is available at: https://github.com/CAMMA-public/Hi-FL
Keywords: Surgical Phase Recognition; Federated Learning; VLM.

1 Introduction

Surgical phase recognition is a fundamental task for next-generation computer-assisted intervention (CAI) systems, enabling context-aware intraoperative assistance and automated workflow documentation [3], [7]. Despite its importance,

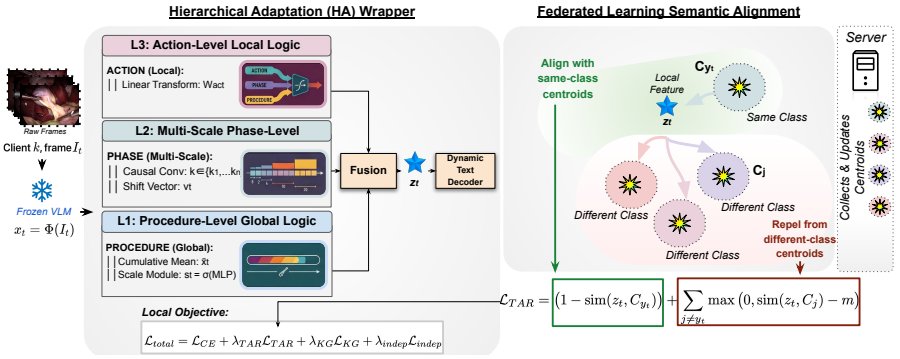


Fig. 1: Hierarchical Adaptation (HA) with a Federated Learning (FL) pipeline. Each branch corresponds to one hierarchical level (L1–L3) and the dynamic text decoder, matching the ablation variants A1–A3 in Table 2.

the development of generalizable models is blocked by the siloed nature of surgical data: strict privacy regulations (e.g., GDPR, HIPAA) prevent the centralized aggregation of diverse datasets necessary to overcome domain shifts [16], [14].

To bridge this gap, recent research has turned to Vision-Language Models (VLMs). While general-purpose models like CLIP provide a baseline for semantic grounding [18], surgical-specific VLMs such as PeskaVLP and HecVL have successfully internalized hierarchical procedural knowledge through large-scale centralized pre-training [26], [25].

However, even these models require center-specific adaptation to remain efficient in diverse clinical settings. Every institution operates within a unique "dialect" defined by specific imaging hardware, different instrument brands, robotic vs. laparoscopic platforms, and surgeon-specific workflows. This field faces *two fundamental bottlenecks*. First, existing surgical VLMs encode procedural hierarchy into static, offline representations. This lack of temporal flexibility prevents the model from adapting to center-specific variations in phase duration or tool-tissue interaction styles. Second, while Federated Learning (FL) provides a framework for decentralized collaboration [2], standard optimization algorithms are "structure-blind" [15], [9]: they treat universal procedural logic and institution-specific noise as a uniform feature set, causing *semantic drift* [27].

In this paper, we argue that effective surgical AI requires *hierarchical disentanglement* during the adaptation of VLMs across federated centers. To this end, we introduce the Hi-FL framework with the following contributions:

- **Hierarchical Adaptation (HA):** A vision-language wrapper that explicitly decomposes surgical reasoning into three structured levels – *Procedure*, *Phase*, and *Action* – enabling semantically grounded adaptation while preserving global procedural invariants.
- **Federated Learning Semantic Alignment:** A novel federated optimization objective designed to address statistical heterogeneity. Class-Conditional

Targeted Alignment and Repulsion (TAR) loss aligns local representations with shared global class semantics while actively repelling cross-institutional domain interference.
– **Extensive Evaluation:** We evaluate our framework across three distinct procedures – Cholecystectomy, Colorectal surgery, and Gastric-Bypass – demonstrating consistent performance gains, particularly in data-scarce (5%–50% per center) and low-shot (16–64 frames) regimes.

2 Related Work

2.1 Hierarchical Surgical Workflow Analysis

Surgical workflow analysis has evolved from Hidden Markov Models to deep architectures that decompose procedures into nested semantic levels [1], [7]. Recent VLMs such as PeskaVLP [25] and HecVL [26] incorporate hierarchical knowledge during pre-training to align video features with procedural hierarchies. Despite these advances, existing hierarchical VLMs operate as centralized, offline backbones with static representations. In contrast, our framework introduces *hierarchical adaptation*, which disentangles surgical reasoning into levels while enabling dynamic, site-aware refinement under decentralized training.

2.2 Adaptation Strategies for Medical Vision-Language Models

Medical VLM adaptation under limited supervision has largely relied on parameter-efficient strategies. Linear probing with text blending (LP+Text) anchors visual prototypes to fixed textual embeddings through learnable multipliers [20], outperforming prompt-learning approaches such as CoOp [30], CoCoOp [29], and KgCoOp [23]. Surgical Phase Anywhere (SPA) [24] proposes a lightweight few-shot framework combining spatial alignment, task-graph priors, and test-time adaptation. While effective, SPA focuses on post-hoc inference-time refinement and does not enforce hierarchical disentanglement during optimization. In contrast, our method embeds hierarchical semantic decomposition directly into federated training.

2.3 Federated Adaptation of Vision-Language Models

Federated Learning (FL) is a standard paradigm for privacy-preserving collaboration across decentralized medical data silos [4], [2]. FedCLIP [10] proposed generalization and personalization for CLIP; F3OCUS uses Neural Tangent Kernels to identify critical VLM layers for local adaptation [19]; FedAdapterProto integrates convolutional adapters with spatial prototypes [17]. Semantic alignment methods such as FedSA [31] and FedTSP [22] leverage language-driven prototypes to mitigate distribution shifts but fail to model the sequential logic of surgical procedures. Our framework addresses these limitations by integrating *hierarchical understanding* directly into federated optimization.

3 Methodology

We present Hi-FL, which integrates *Hierarchical Adaptation* with a *Federated Learning Semantic Alignment* pipeline (Figure 1).

3.1 Problem Formulation

Let $V = \{I_t\}_{t=1}^{T}$ be a surgical video of T frames. Our goal is to predict frame-wise phase labels $\hat{y}_t \in \{1, \ldots, C\}$, causally $\hat{y}_t = f_\Theta(I_{\leq t})$. In the federated setting, K clinical centers hold local datasets $\mathcal{D}_k$. We aim to learn a global model Θ by aggregating local updates $\{\theta_1, \ldots, \theta_K\}$ without centralizing raw video data.

3.2 Hierarchical Adaptation

Hierarchical Adaptation (HA) decomposes surgical logic into three hierarchical levels that operate on VLM visual features $x_t \in \mathbb{R}^D$, producing a refined embedding z_t via hierarchical residual fusion. Each level corresponds to an ablation variant in Section 5.4.

(L1) Procedure-Level (ablated by A2): A causal cumulative mean $\bar{x}_t = \frac{1}{t}\sum_{i=1}^{t} x_i$ captures overall surgical progress. The *Procedure-Scale* module $\mathcal{F}_{proc}(\cdot)$, implemented as a bottleneck MLP, generates a bounded gating vector $s_t \in (0,1)^D$ that emphasizes stage-relevant features: $h_t^{proc} = s_t \odot x_t$.

(L2) Phase-Level (ablated by A1): Phases vary in duration and transition rate. A Multi-Scale Causal Pyramid applies three parallel 1D depth-wise convolutions with kernels $k \in \{k_1, \ldots, k_n\}$, left-padded to enforce strict causality, producing an additive shift vector v_t that captures multi-resolution temporal patterns.

(L3) Action-Level: A learnable projection $a_t = W_{\text{act}} x_t$, with $W_{\text{act}} \in \mathbb{R}^{D\times D}$ initialized as the identity, acts as a feature-space filter. It amplifies motion-sensitive dimensions while suppressing irrelevant VLM features via frame-level loss gradients. The levels combine as: $z_t = x_t + \mathcal{G}(x_t) \odot \big((a_t \odot s_t) + v_t\big)$, where $\mathcal{G}(\cdot)$ is a learnable gate and $\odot$ is element-wise multiplication.

Dynamic Semantic Text Decoder (ablated by A3): A Meta-Net $M(\cdot)$ maps $\bar{x}_t$ to a dynamic shift δ_t applied to text embeddings w_c of phase c: $\hat{w}_{c,t} = \text{Norm}(w_c + \delta_t)$, ensuring textual representations reflect the current procedural context.

Final frame probabilities are computed as the cosine similarity between visual z_t and text $\hat{w}_{c,t}$ embeddings: $P(y_t = c \mid I_t) = \frac{\exp(\tau \cdot \cos(z_t, \hat{w}_{c,t}))}{\sum_j \exp(\tau \cdot \cos(z_t, \hat{w}_{j,t}))}$.

3.3 Federated Learning Semantic Alignment

To handle domain heterogeneity across centers, we introduce the *Class-Conditional Targeted Alignment and Repulsion (TAR)* loss. Each center computes and shares *class-wise visual centroids* $\{C_c\}_{c=1}^{C}$ with the server: $C_c^{(k)} = \frac{1}{|\mathcal{D}_k^c|}\sum_{x\in\mathcal{D}_k^c} z_t$, where

$\mathcal{D}_k^c$ denotes the set of frames with label c at center k. These centroids act as *proxies for canonical class representations*, capturing stable procedural semantics while averaging out local variations.

The TAR loss encourages z_t to align with its class centroid C_{y_t} while repelling other centroids, preserving inter-class separation across institutions:

$$\mathcal{L}_{TAR} = \big(1 - \text{sim}(z_t, C_{y_t})\big) + \sum_{j \neq y_t} \max\big(0, \text{sim}(z_t, C_j) - m\big), \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and m is a margin hyperparameter.

The full local training objective combines classification, hierarchical alignment, and regularization:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{TAR}\mathcal{L}_{TAR} + \lambda_{KG}\mathcal{L}_{KG} + \lambda_{indep}\mathcal{L}_{indep}. \tag{2}$$

– *Knowledge Guard ($\mathcal{L}_{KG}$).* Preserves pre-trained VLM knowledge by minimizing MSE between dynamic text weights $\hat{w}_{c,t}$ and anchor embeddings w_c:

$$\mathcal{L}_{\text{KG}} = \frac{1}{C} \sum_{c=1}^{C} \|\hat{w}_{c,t} - w_c\|_2^2 \tag{3}$$

– *Feature Independence ($\mathcal{L}_{indep}$).* Encourages hierarchical disentanglement by penalizing the absolute cosine similarity between temporally mean-centered Procedure (s_t) and Phase (v_t) vectors, forcing L1 and L2 to find orthogonal solutions:

$$\mathcal{L}_{\text{indep}} = \left|\cos(\tilde{s},\, \tilde{v})\right|, \quad \tilde{s} = s_t - \bar{s}_t, \quad \tilde{v} = v_t - \bar{v}_t, \quad \bar{s}_t = \frac{1}{T}\sum_{t=1}^{T} s_t, \quad \bar{v}_t = \frac{1}{T}\sum_{t=1}^{T} v_t \tag{4}$$

4 Experiments

Datasets. We evaluate on three datasets: Bypass [8] (two-center gastric bypass), a five-center in-house Colorectal dataset using the phase definitions in [5] (32 train; 11 val; 11 test videos in total), and Multicholec [6] (five-center laparoscopic cholecystectomy). Frame-wise surgical annotation is exceptionally costly, requiring expert surgeons to label long procedures frame by frame [13]; consequently, these three datasets represent the practical upper bound of publicly available or institutionally accessible FL benchmarks for surgical phase recognition. Following the splits [8], [6], we vary training data from 5% to 50% (nested subsets) to assess robustness under limited supervision. We also evaluate three supervision levels (16, 32, 64 shots per video), sampled across the full duration and approximately class-balanced.

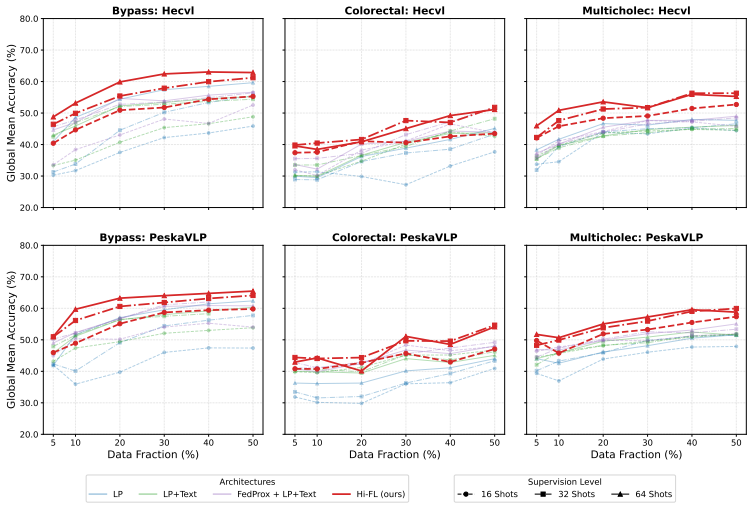


Fig. 2: Performance evaluation of Hi-FL against LP, LP+Text, and FedProx LP+Text across three surgical datasets (5%–50% data, 16–64 shots). Shaded bands denote ±1 std across seeds and centers.

Training Details. Models are trained with AdamW (lr $= 1\times10^{-3}$, weight decay$= 1\times10^{-4}$) and batch size 1, using gradient clipping (max norm 1.0) for stability. Loss weights are set to 0.1 for $\mathcal{L}_{\text{KG}}$, 0.3 for $\mathcal{L}_{\text{TAR}}$, and 0.1 for $\mathcal{L}_{\text{indep}}$, with a 0.4 repulsion margin for $\mathcal{L}_{\text{TAR}}$. Training runs for max 50 FedAvg communication rounds with early stopping [21], exchanging class-wise centroids each round. Results are averaged over three seeds with fixed initialization and reported as *Accuracy* and *Macro-F1* [25], [26].

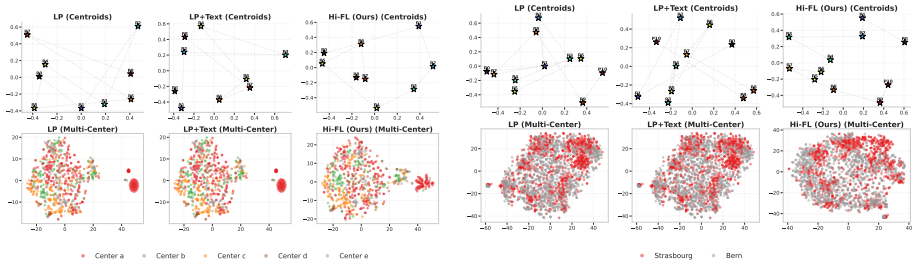
5 Results

5.1 Baseline Comparative Performance

To evaluate Hi-FL, we compare it against three baselines: Linear Probing (LP) with frozen visual features and a linear classifier trained via *FedAvg*; LP+Text [20], which extends LP with class-specific text embeddings under *FedAvg*; and FedProx LP+Text [9], a *FedProx* variant of LP+Text.

Performance analysis across Bypass, Colorectal, and Multicholec (Figure 2) reveals the following trends:

- LP and LP+Text improve with more data (5%–50%) but plateau well below Hi-FL. In *Bypass*, LP+Text (64 shots) reaches ∼0.55 accuracy, while Hi-FL scales towards 0.65.
- Baselines drop sharply under low supervision (16 shots), especially FedProx LP+Text on *Multicholec*, showing lower robustness than Hi-FL.



(a) Colorectal Dataset. (b) Bypass Dataset.

Fig. 3: t-SNE of phase centroids (top) and PCA of multi-center feature distributions (bottom) for LP vs. Hi-FL. Top: phase labels indicate semantic clustering; phases with the same color but different markers denote centroids from different methods. Bottom: colors denote hospital centers; Hi-FL merges center-specific clusters into a single distribution.

– Baseline performance varies with the VLM backbone, unlike the consistent trend of Hi-FL's hierarchical design.

Result Variability and the Role of Federation. The observed variability in results arises from two sources: (i) genuine multi-center heterogeneity – for example, in Bypass, the omentum division phase is routinely performed in Strasbourg but not in Bern, creating a structural distributional mismatch – and (ii) the data-scarce regime (5% data, 16-shot), where small differences in sampled frames affect optimization. Federation consistently acts as a stabilizer: removing it (local training per center, no aggregation) collapses accuracy and inflates variance across all datasets, while Hi-FL achieves the highest accuracy with the lowest inter-run standard deviation. This confirms that the federated aggregation step, combined with the TAR loss and hierarchical disentanglement, is central to stabilizing training under heterogeneous, low-data conditions.

5.2 Alignment and Multi-Center Robustness

We assess Hi-FL's alignment versus LP and LP+Text using t-SNE [11] and PCA [12] on phase centroids and multi-center feature distributions (Figure 3), focusing on Colorectal and Bypass due to their high inter-center variability.

For the Colorectal dataset (Fig. 3a), LP centroids for P2 (Dissection) and P4 (Anastomosis) overlap, while Hi-FL organizes phases clearly, with P7 and P8 forming tight clusters. LP distributions are fragmented by hospital site; Hi-FL produces cohesive overlap across all five centers. For the Bypass dataset (Fig. 3b), LP centroids compress overlapping phases P1 and P7, whereas Hi-FL expands semantic distances while maintaining phase order. LP forms two separate "islands" for Strasbourg and Bern; Hi-FL merges them into a single cloud.

Table 1: Global Mean Accuracy/F1 (%) ± Mean Inter-site std (5%, 16-shot).

Dataset	CoCoop[29]	FedCLIP[10]	FedPrompt[28]	FedSA[31]	LP+Text[20]	**Hi-FL**
Colorectal	$40.3/28.1_{\pm 2.0}$	$32.1/19.9_{\pm 2.8}$	$29.7/21.8_{\pm 2.9}$	$33.9/24.6_{\pm 2.4}$	$41.5/29.1_{\pm 1.9}$	$\mathbf{42.5}/\mathbf{31.3}_{\pm \mathbf{2.2}}$
Multicholec	$47.3/38.5_{\pm 3.9}$	$42.0/28.1_{\pm 3.9}$	$39.7/30.8_{\pm 3.7}$	$41.2/34.7_{\pm 4.6}$	$46.3/38.0_{\pm 3.9}$	$\mathbf{51.9}/\mathbf{40.5}_{\pm \mathbf{4.1}}$
Bypass	$44.6/\mathbf{26.1}_{\pm \mathbf{5.9}}$	$46.6/21.6_{\pm 6.3}$	$46.4/22.3_{\pm 5.8}$	$43.1/24.0_{\pm 5.9}$	$45.0/21.8_{\pm 5.3}$	$\mathbf{47.9}/22.8_{\pm 6.1}$

5.3 Federated Learning Adaptation

Table 1 shows that Hi-FL consistently outperforms prior FL adaptation methods across all datasets under the strict data regime (5%, 16-shot), achieving the highest mean accuracy and F1-scores in most cases while also keeping lower inter-site variability.

5.4 Ablation and Sensitivity Analysis

On 5% Colorectal and Multicholec data (16-shot), Hi-FL with PeskaVLP shows the Dynamic Text Decoder (A3) is essential: removing it drops accuracy and markedly increases variance, demonstrating that static prompts cannot bridge the surgical FL domain gap. The temporal pyramid (A1) stabilizes F1, and global context (A2) slightly reduces it. As shown in Figure 4, Hi-FL produces more coherent, temporally aligned predictions than the baselines.

Table 2: Ablation (A) and Sensitivity (S) at 5% data and 16-shot.

	Colorectal		**Multicholec**	
Variant	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**
Hi-FL	$\mathbf{42.5}_{\pm \mathbf{2.1}}$	$\mathbf{31.3}_{\pm \mathbf{2.0}}$	$\mathbf{51.9}_{\pm \mathbf{3.2}}$	$\mathbf{40.5}_{\pm \mathbf{5.0}}$
A1: No Temp	$42.4_{\pm 2.2}$	$28.8_{\pm 1.8}$	$48.9_{\pm 2.9}$	$37.9_{\pm 5.4}$
A2: No Glob	$43.4_{\pm 2.5}$	$30.1_{\pm 2.4}$	$49.2_{\pm 3.1}$	$37.2_{\pm 5.6}$
A3: Static	$35.9_{\pm 7.7}$	$23.5_{\pm 5.1}$	$43.6_{\pm 5.0}$	$34.2_{\pm 4.2}$
S1: Weak	$42.5_{\pm 3.2}$	$29.7_{\pm 3.2}$	$50.2_{\pm 3.0}$	$38.6_{\pm 4.5}$
S2: M=0.1	$41.2_{\pm 4.2}$	$28.3_{\pm 3.1}$	$49.2_{\pm 2.7}$	$38.1_{\pm 5.0}$
S3: M=0.7	$41.8_{\pm 4.1}$	$28.6_{\pm 3.6}$	$49.7_{\pm 2.8}$	$38.0_{\pm 5.3}$

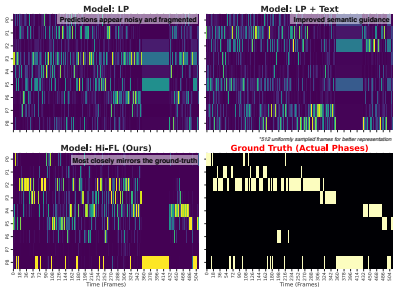


Fig. 4: Random Colorectal video: Hi-FL, LP, LP+Text phase predictions vs. ground truth. Hi-FL yields smoother, semantically consistent transitions.

Lowering the $\mathcal{L}_{KG}$ weight (S1) raises variance or lowers accuracy, showing the need to anchor to pre-trained weights. Geometric Margin analysis (S2, S3) finds $m = 0.4$ optimal: looser margins merge phases, stricter ones fragment features. Sensitivity across a $10\times$ $\mathcal{L}_{KG}$ weight range ($w_{kg} = 0.01$–0.1) and margin range 0.1–0.7 confirms robustness, with <1pp accuracy variation on Colorectal and <2pp on Multicholec/Bypass.

6 Discussion

Hi-FL shows consistent gains across procedures and federated settings, but has several limitations. First, dense frame-wise annotation requires expert labeling of long procedures [13], making the 5%–50% label regime a realistic deployment constraint. Second, the framework primarily assumes *phase-distributional* heterogeneity; settings with substantial imaging hardware differences may require additional domain adaptation. Third, Hi-FL depends on expert labels; future work will explore weakly and self-supervised extensions.

7 Conclusions

Hi-FL hierarchically adapts pre-trained VLMs for federated surgical datasets, aligning cross-center features while disentangling three-level representations. It outperforms baselines by up to 10% in 5%–50% data regimes, preserves semantic phase structure, and offers a practical framework for multi-center deployment.

Acknowledgments. This work was supported by French state funds managed within the Plan Investissements d'Avenir by the ANR under references ANR-22-FAI1-0001 (project DAIOR), ANR-10-IAHU-02 (IHU Strasbourg) and ANR-23-IACL-0004 (AI Cluster Grand Est ENACT).

Disclosure of Interests. Pietro Mascagni and Nicolas Padoy are co-founders and shareholders of Scialytics. The remaining authors have no competing interests to declare.

References

1. Ayobi, N., Rodríguez, S., Pérez, A., Hernández, I., Aparicio, N., Dessevres, E., Peña, S., Santander, J., Caicedo, J.I., Fernández, N., et al.: Pixel-wise recognition for holistic surgical scene understanding. Medical Image Analysis p. 103726 (2025)
2. Ciupek, D., Malawski, M., Pieciak, T.: Federated learning: A new frontier in the exploration of multi-institutional medical imaging data. arXiv preprint arXiv:2503.20107 (2025)
3. Garrow, C.R., Kowalewski, K.F., Li, L., Wagner, M., Schmidt, M.W., Engelhardt, S., Hashimoto, D.A., Kenngott, H.G., Bodenstedt, S., Speidel, S., et al.: Machine learning for surgical phase recognition: a systematic review. Annals of surgery **273**(4), 684–693 (2021)
4. Guan, H., Yap, P.T., Bozoki, A., Liu, M.: Federated learning for medical image analysis: A survey. Pattern recognition **151**, 110424 (2024)
5. Jain, P.P., Mascagni, P., Massimiani, G., Banik, N., Goglia, M., Arboit, L., Baby, B., Balla, A., Baldari, L., Silecchia, G., et al.: Expert consensus-based video-based assessment tool for workflow analysis in minimally invasive colorectal surgery: Development and validation of coloworkflow. arXiv preprint arXiv:2511.10766 (2025)

6. Kassem, H., Alapatt, D., Mascagni, P., Karargyris, A., Padoy, N.: Federated cycling (fedcy): Semi-supervised federated learning of surgical phases. IEEE transactions on medical imaging **42**(7), 1920–1931 (2022)
7. Khan, U., Nawaz, U., Qayyum, A., Ashraf, S., Xie, Y., Khan, M.H., Bilal, M., Qadir, J.: Surgical scene understanding in the era of foundation ai models: A comprehensive review. arXiv preprint arXiv:2502.14886 (2025)
8. Lavanchy, J.L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B.P., Nett, P.C., Marescaux, J., Mutter, D., Padoy, N.: Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. International journal of computer assisted radiology and surgery **19**(11), 2249–2257 (2024)
9. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020)
10. Lu, W., Hu, X., Wang, J., Xie, X.: Fedclip: Fast generalization and personalization for clip in federated learning. arXiv preprint arXiv:2302.13485 (2023)
11. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
12. Maćkiewicz, A., Ratajczak, W.: Principal components analysis (pca). Computers & Geosciences **19**(3), 303–342 (1993)
13. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al.: Surgical data science–from concepts toward clinical translation. Medical image analysis **76**, 102306 (2022)
14. Mansour, E., Srinivas, K., Hose, K.: Federated data science to break down silos [vision]. ACM SIGMOD Record **50**(4), 16–22 (2022)
15. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. Pmlr (2017)
16. Moore, W., Frye, S.: Review of hipaa, part 1: history, protected health information, and privacy and security rules. Journal of nuclear medicine technology **47**(4), 269–272 (2019)
17. Ofosu Mensah, S., Djoumessi, K., Berens, P.: Prototype-guided and lightweight adapters for inherent interpretation and generalisation in federated learning. In: MICCAI. pp. 464–473. Springer (2025)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
19. Saha, P., Wagner, F., Mishra, D., Peng, C., Thakur, A., Clifton, D.A., Kamnitsas, K., Noble, J.A.: F^ 3ocus-federated finetuning of vision-language foundation models with optimal client layer updating strategy via multi-objective meta-heuristics. In: CVPR. pp. 20006–20017 (2025)
20. Shakeri, F., Huang, Y., Silva-Rodríguez, J., Bahig, H., Tang, A., Dolz, J., Ben Ayed, I.: Few-shot adaptation of medical vision-language models. In: MICCAI. pp. 553–563. Springer (2024)
21. Sun, T., Li, D., Wang, B.: Decentralized federated averaging. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4289–4301 (2022)
22. Wu, X., Niu, J., Liu, X., Zhu, G., Zhang, J., Tang, S.: Enhancing visual representation with textual semantics: Textual semantics-powered prototypes for heterogeneous federated learning. arXiv preprint arXiv:2503.13543 (2025)
23. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: CVPR. pp. 6757–6767 (2023)

24. Yuan, K., Chen, T., Li, S., Lavanchy, J.L., Heiliger, C., Özsoy, E., Huang, Y., Bai, L., Navab, N., Srivastav, V., et al.: Recognizing surgical phases anywhere: Few-shot test-time adaptation and task-graph guided refinement. In: MICCAI. pp. 467–477. Springer (2025)
25. Yuan, K., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. Advances in Neural Information Processing Systems **37**, 122952–122983 (2024)
26. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: MICCAI. pp. 306–316. Springer (2024)
27. Zeng, S., Guo, P., Wang, S., Wang, J., Zhou, Y., Qu, L.: Tackling data heterogeneity in federated learning via loss decomposition. In: MICCAI. pp. 707–717. Springer (2024)
28. Zhao, H., Du, W., Li, F., Li, P., Liu, G.: Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In: ICASSP. pp. 1–5. IEEE (2023)
29. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
30. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022)
31. Zhou, Y., Qu, X., You, C., Zhou, J., Tang, J., Zheng, X., Cai, C., Wu, Y.: Fedsa: A unified representation learning via semantic anchors for prototype-based federated learning. In: AAAI. vol. 39, pp. 23009–23017 (2025)