MICCAI2025 PAPER HIGHLIGHT

MICCAI2025 Daejeon REPUBLIC OF KOREA

Adaptation of Multi-modal Representation Models for Multi-task Surgical Computer Vision

Soham Walimbe[1], Britty Baby[1,2], Vinkle Srivastav[1,2], and Nicolas Padoy[1,2]

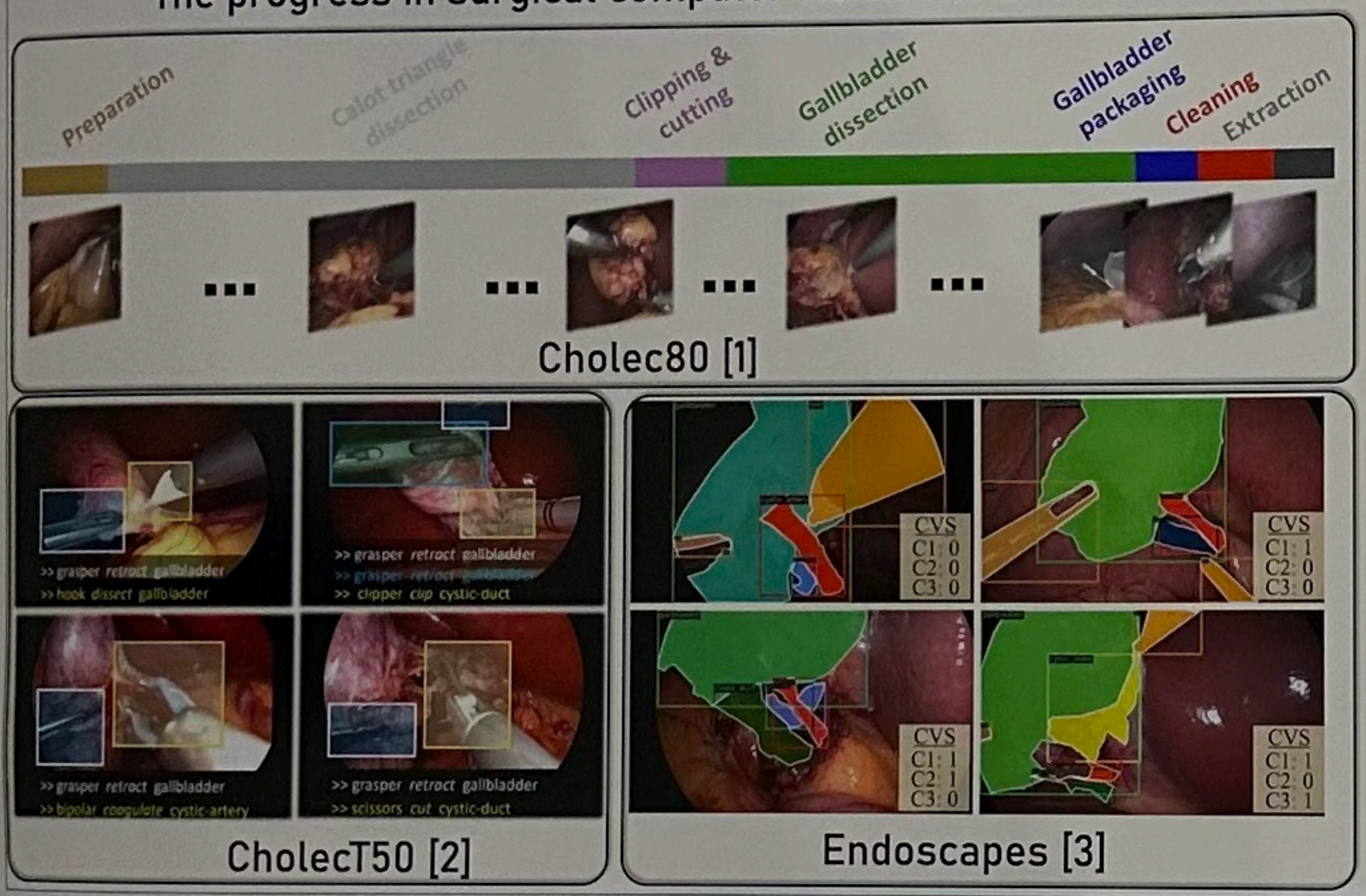
1 University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France
2 Institute of Image-Guided Surgery, IHU Strasbourg, Strasbourg, France

Abstract

Surgical AI often involves multiple tasks within a single procedure, like phase recognition or assessing the CVS in laparoscopic cholecystectomy. Traditional models, built for one task at a time, lack flexibility, requiring a separate model for each. To address this, we introduce MML-SurgAdapt, a unified multi-task framework with Vision-Language Models (VLMs), specifically CLIP, to handle diverse surgical tasks through natural language supervision. A key challenge in multi-task learning is the presence of partial annotations when integrating different tasks. To overcome this, we employ Single Positive Multi-Label (SPML) learning, which traditionally reduces annotation burden by training models with only one positive label per instance. Our framework extends this approach to integrate data from multiple surgical tasks within a single procedure, enabling effective learning despite incomplete or noisy annotations. We demonstrate the effectiveness of our model on a combined dataset consisting of Cholec80, Endoscapes2023, and CholecT50, utilizing custom prompts. Extensive evaluation shows that MML-SurgAdapt performs comparably to task-specific benchmarks, with the added advantage of handling noisy annotations. To our knowledge, this is the first application of SPML to integrate data from multiple surgical tasks, presenting a novel and generalizable solution for multi-task learning in surgical CV.

Motivation

The progress in surgical computer vision has been focused on proposing diverse tasks leading to fragmented task-specific models


Preparation
Calot triangle dissection
Clipping & cutting
Gallbladder dissection
Gallbladder packaging
Cleaning
Extraction
>> grasper retract gallbladder
>> hook dissect gallbladder
>> grasper retract gallbladder
>> clipper clip cystic-duct
>> grasper retract gallbladder
>> bipolar coagulate cystic-artery
>> grasper retract gallbladder
>> scissors cut cystic-duct
CVS C1: 0 C2: 0 C3: 0
CVS C1: 1 C2: 0 C3: 0
CVS C1: 1 C2: 1 C3: 0
CVS C1: 1 C2: 0 C3: 1

Cholec80 [1]

CholecT50 [2]

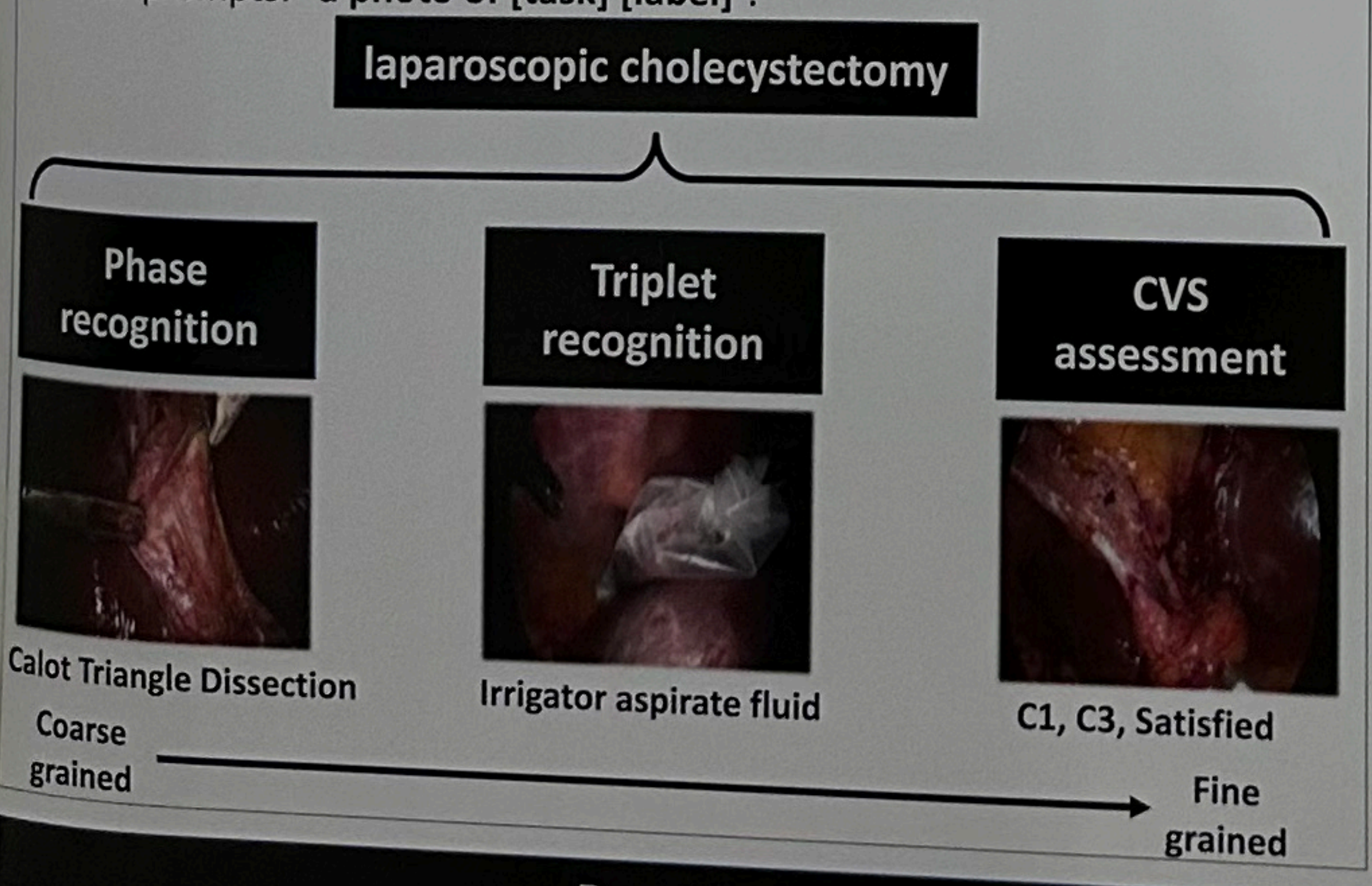
Endoscapes [3]

- **Task-specific dataset curation:** Existing surgical video datasets—**Cholec80**, **CholecT50**, and **Endoscapes**—have each been curated with distinct objectives, resulting in specialized, task-specific models.
- **Heterogeneous problem formulations:**
  - *Cholec80* focuses on **surgical workflow recognition**, formulated as a **multi-class classification problem**.
  - *CholecT50* addresses **fine-grained action triplet recognition**, framed as a **multi-label classification task**.
  - *Endoscapes* focuses on **anatomy-driven critical view of safety** recognition, also formulated as a **multi-label classification task**.

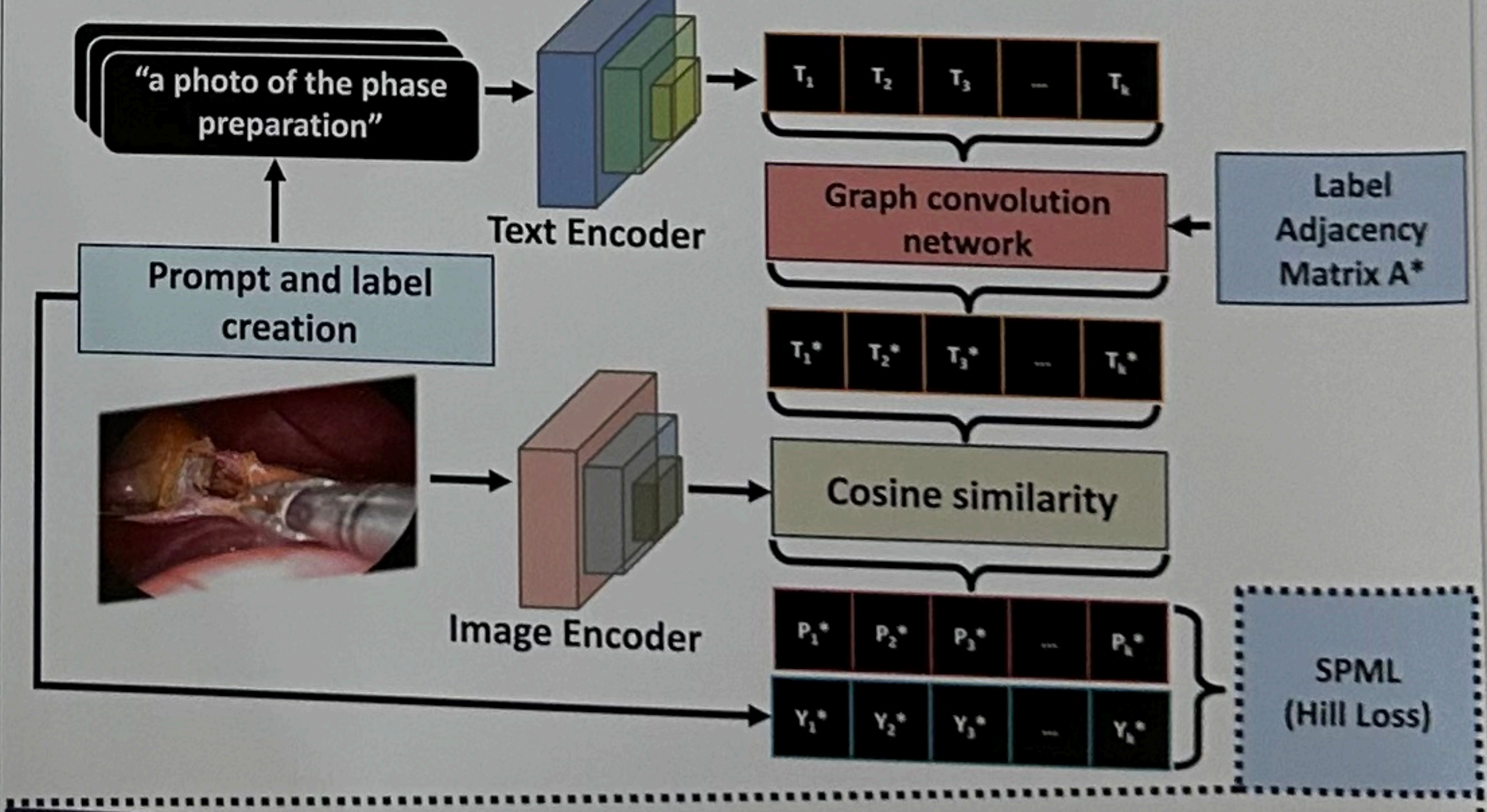
How can partially annotated datasets be effectively combined to overcome dataset boundaries and enable the development of unified and generalizable models across diverse surgical tasks?

Dataset

- Unified labels **from Cholec80 (7 classes)**, **CholecT50 (100 classes)** and **Endoscapes (3 classes)** into a 110-label multi-task space.
- Each image keeps **one** positive label from its **task-specific annotation**, while labels from other tasks treated as **negatives**.
- Label prompts: "**a photo of [task] [label]**".



Method: MML-SurgAdapt

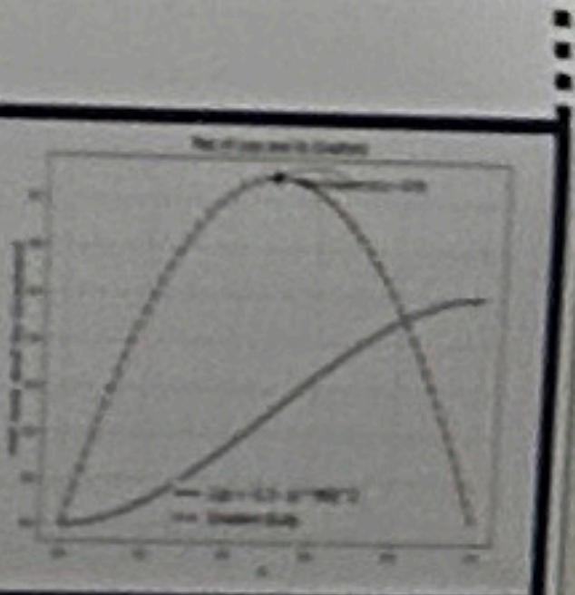


- Reweights negatives by **down-weighting** high confidence predictions to reduce false negatives
- Applies **semi-hard positive mining** to focus learning on informative positives.

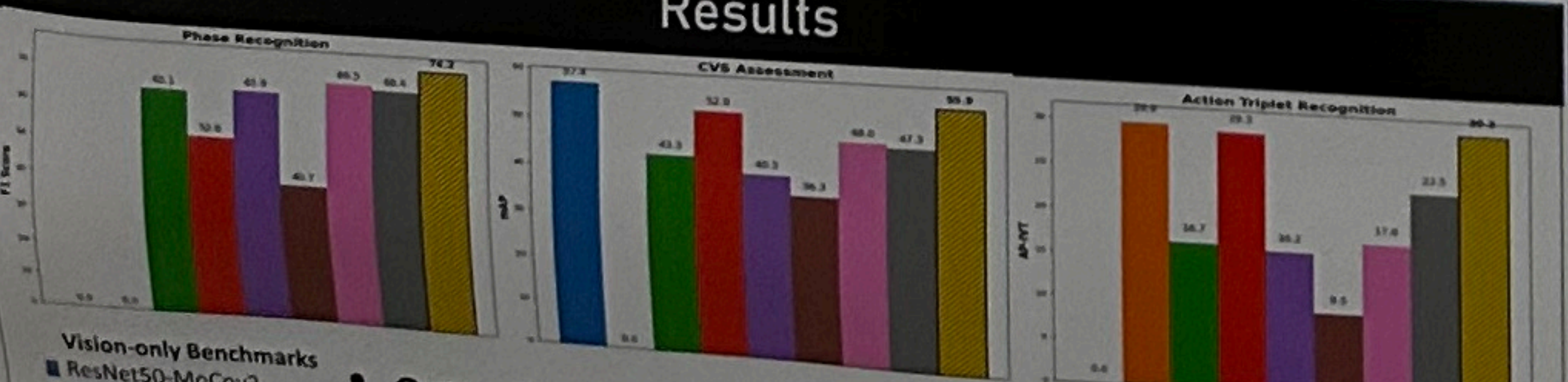
$$p_{km} = \sigma\left(\frac{s_k}{\tau} - m\right)$$

$$\text{loss}^{+}_{\text{Hill}} = -(1 - p_{km})^{\gamma} \cdot \log(p_{km})$$

$$\text{loss}^{-}_{\text{Hill}} = -(\lambda - p_k) \cdot p_k^2$$



Results



- **Outperforms** task-specific and SPML baselines, while **competitive** with vision-only benchmarks on complete annotations.
- **Limitations** include single-procedure evaluation and class imbalance; future work involves broader **surgical validation** and **domain-specific** pretraining.
- GCN provided minimal gain; better adaptation may require surgical domain pretraining

Conclusion

- **MML-SurgAdapt** unifies diverse surgical tasks with label efficient SPML learning.
- With SPML learning, **we tackle the challenges of combining datasets** for interrelated tasks, mitigating false negatives, and enabling label-efficient learning from partial annotations.
- Achieves comparable or superior performance to task-specific and SPML benchmarks.

References

[1] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. TMI (2016)
[2] Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. MedIA (2022)
[3] Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., ... & Padoy, N. (2023). Latent graph representations for critical view of safety assessment. TMI (2023)