fine-CLIP: Enhancing Zero-Shot Fine-Grained Surgical Action Recognition with Vision-Language Models

Saurav Sharma[a,b,1], Didier Mutter[b,c], Nicolas Padoy[a,b]

[a]University of Strasbourg, CNRS, INSERM, ICube, UMR7357, France
[b]IHU Strasbourg, Strasbourg, France
[c]University Hospital of Strasbourg, France

While vision-language models like CLIP have advanced zero-shot surgical phase recognition, they struggle with fine-grained surgical activities, especially action triplets. This limitation arises because current CLIP formulations rely on global image features, which overlook the fine-grained semantics and contextual details crucial for complex tasks like zero-shot triplet recognition. Furthermore, these models do not explore the hierarchical structure inherent in triplets, reducing their ability to generalize to novel triplets. To address these challenges, we propose fine-CLIP, which learns object-centric features and leverages the hierarchy in triplet formulation. Our approach integrates three components: hierarchical prompt modeling to capture shared semantics, LoRA-based vision backbone adaptation for enhanced feature extraction, and a graph-based condensation strategy that groups similar patch features into meaningful object clusters. Since triplet classification is a challenging task, we introduce an alternative yet meaningful base-to-novel generalization benchmark with two settings on the CholecT50 dataset: Unseen-Target, assessing adaptability to triplets with novel anatomical structures, and Unseen-Instrument-Verb, where models need to generalize to novel instrument-verb interactions. fine-CLIP shows significant improvements in F1 and mAP, enhancing zero-shot recognition of novel surgical triplets.

Keywords: Vision-Language Model, Surgical Action Triplets, Zero-Shot Recognition.

1. Introduction

Analyzing surgical workflows is vital in surgical data science [19], facilitating the development of context-aware decision systems that can aid surgeons in the operating room (OR). These systems utilize streaming video data to create comprehensive solutions for enhanced surgical decision-making. They focus on identifying surgical activities at various granularities, such as phases [16], steps [9], and surgical action triplets [11, 13, 14]. Vision-language (VL) models, such as CLIP [12], represent a significant shift from traditional vision-only models, enabling robust zero-shot image classification. This capability is especially valuable in the context of surgical procedures, where it allows for data-efficient learning despite limited labeled data and high annotation costs. By leveraging VL models, adaptable systems can interpret complex surgical triplets without the need for extensive retraining. These models are trained on large image-text datasets using contrastive learning, enabling them to encode open-vocabulary concepts in a shared semantic space. Building on this foundation, SurgVLP [20], trained on surgical image-text data, surpasses CLIP in surgical phase and triplet recognition. Subsequent works [21, 22, 6] enhanced performance using CLIP-style contrastive learning with ResNet-50 and Bio-ClinicalBERT encoders, representing various hierarchical levels in surgical videos and improving zero-shot phase recogni-

tion. VidLPRO [4] further advanced multi-modal representation by introducing multiple pretraining objectives alongside contrastive learning.

However, zero-shot performance on surgical action triplets [11] remains challenging. These triplets, defined as ⟨*instrument, verb, target*⟩, represent fine-grained tool-tissue interactions in surgical scenes, requiring models to disentangle and associate multiple concepts. Current approaches like CLIP and SurgVLP face two key limitations: (1) they rely on flattened features from ViT-B's [1] CLS token, failing to capture fine-grained interactions, and (2) their zero-shot evaluation doesn't effectively test generalization. A more practical approach would be to fine-tune on a subset of triplets and evaluate on unseen ones, better reflecting real-world adaptability to new surgical interactions.

Fine-tuning large-scale VL models is a challenging task, prompting recent research [24, 23] to develop efficient strategies that preserve CLIP's zero-shot capabilities while boosting performance on new tasks. These strategies are divided into two main categories: **prompt-learning** and **adaptation**. **Prompt-learning** methods, such as CoOp [24] and Dual-CoOp [15], involve learning prompt vectors that integrate with class prompts, allowing models to adapt to downstream tasks without altering the visual and text encoders. MaPLe [8] adds multi-modal learnable prompts to both text and vision encoders in CLIP. **Adaptation** methods, like CLIP-Adapter [3], introduce a learnable projection to the vision encoder's features, keeping other weights frozen. LoRA [23, 5] employs rank decomposition matrices within transformer layers for im-
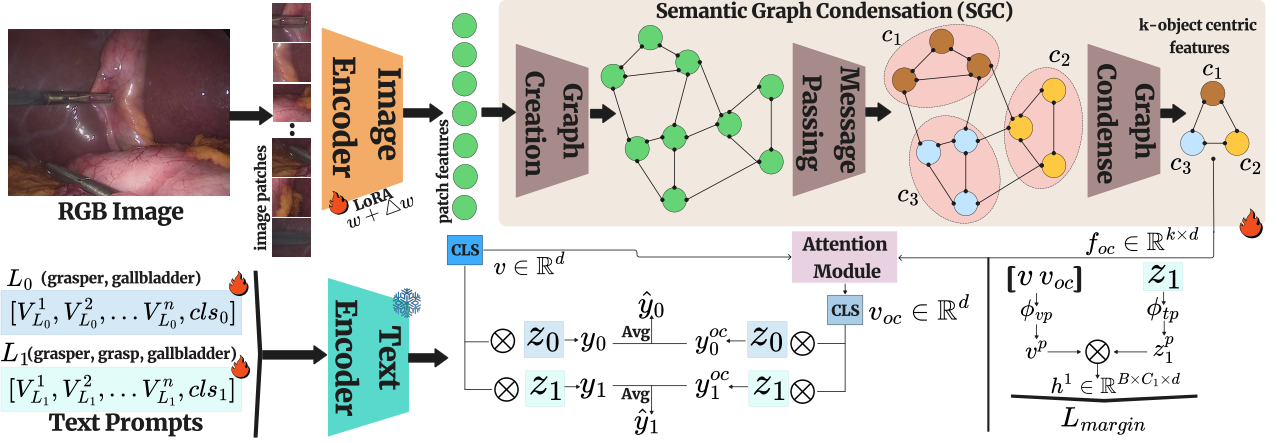
Fig. 1. Model Overview: fine-CLIP tunes the vision backbone using LoRA [23] and extracts object-centric features via Semantic Graph Condensation (SGC). Hierarchical prompts generate two-level text embeddings, while object features enhance image features through attention. A hierarchical margin loss on the combined object-aware and levelwise logits guides the final prediction.

proved task adaptation. Despite these advancements, current methods for zero-shot triplet recognition often fall short, failing to capture the rich context of surgical scenes or utilize the hierarchical structure in triplet annotations. Addressing these limitations, we propose two research questions: (1) How can CLIP's patch features be used to develop object-centric features that enhance fine-grained detail recognition and improve zero-shot triplet performance? (2) Can we exploit the hierarchical relationships in triplet labels for better text and visual feature alignment?

To address these research questions, we introduce **fine-CLIP**, which adapts both text and vision backbones. It integrates learnable prompt vectors in the text encoder, similar to CoOp [24], and employs LoRA [23, 5] style low-rank decomposition in the vision encoder. We observe that patch features from CLIP's vision transformer attention modules contain rich object semantics [2], enhancing global image features. Leveraging these, we develop **SGC** (Semantic Graph Condensation), which constructs a patch affinity graph, performs message passing, and applies graph condensation to cluster similar patch features into object-centric features. These features, combined with global image features through attention mechanisms, improve triplet understanding and boost zero-shot performance. Surgical action triplets exhibit inherent hierarchy; for example, ⟨*grasper, gallbladder*⟩ includes multiple triplets with verbs like *grasp* and *retract*. We leverage this by designing learnable prompts at both ⟨*instrument, target*⟩ (root) and ⟨*instrument, verb, target*⟩ (leaf) levels, capturing broad interactions and specific actions. A max margin loss [18] aligns features with shared parents, enabling hierarchical knowledge transfer for zero-shot classification.

We evaluate fineCLIP's zero-shot generalization across two settings: **unseen target** (UT), where the model encounters new target tissues (e.g., trained on *gallbladder*, tested on *cystic-artery*), and **unseen instrument & verb** (UIV), involving novel instrument-verb combinations. In both settings, our method outperforms existing approaches, including SurgVLP [20] and its variants.

2. Methodology

We introduce **fine-CLIP**, a zero-shot triplet classification model that builds upon CLIP [12]. CLIP employs dual encoders: an image encoder Φ_{img} based on Vision Transformer (ViT) [1] and a text encoder Φ_{text}, each comprising M transformer layers. For visual processing, CLIP divides an image $I \in \mathbb{R}^{C \times H \times W}$ into N patches, adds a class token p_{cls}, and projects them to $\tilde{I} = \{p_{\text{cls}}, p_1, \ldots, p_N\}$, creating a visual embedding $v = \Phi_{\text{img}}(\tilde{I}) \in \mathbb{R}^d$. For text, CLIP encodes a label y using the template "a photo of a {class label}" into $\tilde{T} = \{s_{\text{start}}, w_1, \ldots, w_L, l_y, s_{\text{end}}\}$, with L as the prompt length, yielding a text embedding $t = \Phi_{\text{text}}(\tilde{T}) \in \mathbb{R}^d$.

2.1. Hierarchical Prompt Learning:

Recent work, such as CoOp [24], shows that soft prompting can effectively adapt text encoders with minimal overhead. Given the complexity of surgical triplets, a hierarchical prompting approach captures multi-level relationships for zero-shot classification. This hierarchy includes level L_1 for fine-grained triplet interactions ⟨*instrument, verb, target*⟩ and level L_0 for coarse-grained instrument-tissue relationships ⟨*instrument, target*⟩. The text encoder Φ_{text} processes $t_0 = \{V_{L_0}^1, \ldots, V_{L_0}^n, cls_0\}$ and $t_1 = \{V_{L_1}^1, \ldots, V_{L_1}^n, cls_1\}$ into embeddings $z_0, z_1 \in \mathbb{R}^d$, where cls_0 and cls_1 are class embeddings for levels L_0 and L_1, respectively, with n context tokens (Figure 1).

2.2. Low-Rank Adaptation:

To efficiently adapt large-scale models without full finetuning, we employ Low-Rank Adaptation (LoRA) [5]. LoRA approximates weight updates through low-rank decomposition using matrices $\mathbf{A} \in \mathbb{R}^{r \times d}$ and $\mathbf{B} \in \mathbb{R}^{d \times r}$, where rank $r \ll d$ captures essential parameter updates. For each weight matrix $\mathbf{W}$, the update is applied as $\mathbf{W} + \alpha \mathbf{BA}$, with α as a scaling factor. Following CLIP-LoRA [23], we adapt the attention modules' query, key, and value matrices using low-rank decomposition with rank $r = 2$.

2.3. Learning Object-Centric Features:

CLIP's ViT-B attention key features naturally encode object-level cues [2]. Building on this, our Semantic Graph Condensation (**SGC**) module learns object-centric features without instance supervision. SGC constructs an undirected graph $\mathcal{G}$ from patch features of the jth layer attention key using pairwise cosine similarities between patches. Node features are enriched through Graph Attention Networks (GAT) [17] to aggregate contextual semantics from neighboring patches. Following graph condensation [7], a softmax-normalized assignment matrix maps nodes to k clusters, producing object-centric features $f_{oc} \in \mathbb{R}^{k \times d}$ through matrix multiplication.

2.4. Hierarchical Feature Alignment:

Given object-centric features f_{oc}, visual embedding v, and hierarchical text embeddings z_0 and z_1, we compute logits in two stages. First, following CLIP's approach, we calculate image logits by finding the similarity between L2-normalized image features and text embeddings: $y_1 = \tau(v \cdot z_1^T)$ and $y_0 = \tau(v \cdot z_0^T)$, where τ is a scaling factor. We then enrich image features with object-centric information by computing attention weights: $\alpha = \text{softmax}\left(\frac{v \cdot f_{oc}^T}{\sqrt{d}}\right)$. These weights yield object-aware features $v_{oc} = \alpha f_{oc}$, used to compute another set of logits: $y_1^{oc} = \tau(v_{oc} \cdot z_1^T)$ and $y_0^{oc} = \tau(v_{oc} \cdot z_0^T)$. The final logits are averaged: $\hat{y}_i = \frac{1}{2}(y_i + y_i^{oc})$ for $i \in \{0, 1\}$.

2.5. Optimization:

We apply binary cross entropy loss between predicted logits $\hat{y}_i$ and ground truth labels y_i at each level $i \in \{0, 1\}$ (Equation 1):

$$L_i = \sum_{c=1}^{C_i} \frac{-1}{N}\left(y_{i,c} \log\left(\sigma(\hat{y}_{i,c})\right) + (1 - y_{i,c})\log\left(1 - \sigma(\hat{y}_{i,c})\right)\right),$$
(1)

where C_i refers to the number of classes at level i, $y_{i,c}$ and $\hat{y}_{i,c}$ denote ground truth and predicted labels for class c respectively, and σ is the sigmoid function. The levels L_0 and L_1 exhibit a natural hierarchical relationship, where fine-grained triplets ⟨instrument, verb, target⟩ are subsets of coarse-grained pairs ⟨instrument, target⟩. To leverage this hierarchy and enforce hierarchical constraints, we introduce a max-margin loss operating on level L_1 embeddings. For triplets sharing the same parent (e.g., triplets with same ⟨instrument, target⟩ but different verbs), their embeddings should be closer compared to triplets with different parents. We concatenate image features v with object-aware features v_{oc} and project them to $\mathbb{R}^{\hat{d}}$ using ϕ_{vp} to obtain v^p. Similarly, text embeddings z_1 are projected through ϕ_{tp} to obtain z_1^p. For level L_1, we compute class embeddings $h^1 = v^p \cdot z_1^p$, $h^1 \in \mathbb{R}^{B \times C_1 \times d}$. The hierarchical loss (Equation 2) is defined as:

$$L_{margin} = \frac{1}{|P|} \sum_{(i,j) \in P} \sum_{k \notin N_{ij}} \max(0, m + d(h_i^1, h_j^1) - d(h_i^1, h_k^1)), \quad (2)$$

where P is the set of positive pairs sharing the same parent within a batch, N_{ij} contains batch-sampled negative triplets

not sharing parents with i or j, $d(\cdot, \cdot)$ is cosine distance, and m is the margin. The final loss combines hierarchical and classification objectives (Equation 3):

$$L_{total} = \alpha_0 \times L_0 + \alpha_1 \times L_1 + \alpha_h \times L_{margin}, \quad (3)$$

where α_0, α_1, and α_h are the respective loss weights.

3. Experiments

3.1. Base-to-Novel Generalization:

We evaluate zero-shot generalization by splitting the 100 triplet classes into base and novel sets. The model is trained solely on the base set and then evaluated on the novel set without additional training. In the **Unseen-Target** (UT) setting, we test generalization to new anatomical targets: the base set (36 triplets) includes interactions with *gallbladder*, *cystic duct*, *cystic pedicle*, *liver*, *omentum*, and *blood vessel*, while the novel set (18 triplets) contains *abdominal wall cavity*, *cystic artery*, *cystic plate*, *fluid*, *gut*, *peritoneum*, and *specimen bag*. In the **Unseen-Instrument-Verb** (UIV) setting, we evaluate generalization to new surgical interactions: the base set (28 triplets) includes instrument-verb pairs like *bipolar-coagulate/dissect* and *grasper-dissect/grasp*, while the novel set (21 triplets) features previously unseen combinations such as *irrigator-aspirate/irrigate*, *grasper-retract/pack*, and *scissors-cut*.

3.2. Dataset and Evaluation Metrics:

We evaluate on the CholecT50 [11] dataset following RDV splits [10]. To prevent data leakage, we exclude images containing both base and novel triplets. In the UT setting, we have ~45K, ~3.5K, and ~12K frames for base training, validation, and test sets, respectively, plus ~0.7K and ~1.9K frames for novel validation and test sets. In the UIV setting, we obtain ~18K, ~1.7K, and ~4.3K frames for base training, validation, and test sets, along with ~1K and ~3.7K frames for novel validation and test sets. We measure performance using mean average precision (mAP) and F1-Score@3, and report base, novel, and their harmonic mean (HM) scores on the test set.

3.3. Implementation Details:

We implement fine-CLIP in PyTorch and train on a single NVIDIA-A100 GPU. We resize input frames to 224×224 and apply RandAugment. Using ViT-B/16 as the image encoder, we train for 30 epochs with AdamW (learning rate: $1e^{-3}$, weight decay: $1e^{-2}$) optimizer and a batch size of 32. The text encoder is the default Transformer used in CLIP [12]. We use $n = 8$ context tokens and set a margin $m = 0.7$ in L_{margin}. We set $d = 512$ and $\tilde{d} = 256$, using 1-layer MLPs for ϕ_{vp} and ϕ_{tp}. SGC uses one GAT [17] layer for message passing and condensation. We tune hyperparameters on the validation set, setting all loss weights (α_0, α_1, α_h) to 1, while baselines use default settings.

Table 1. Zero-shot triplet classification performance (F1@3 and AP in %) on UT setting (P: prompt learning, A: adaptation, ZS: zero-shot, FT: Fine-tuning).

Method	#Trainable Params(MB)	Category	Base		Novel		HM	
			F1@3	mAP	F1@3	mAP	F1@3	mAP
CLIP [12]	0	ZS	0.73	5.64	9.2	5.85	1.34	5.74
CLIP [12]	86.192	FT	57.36	19.29	24.12	18.05	33.96	18.64
CoOp [24]	0.004	P	50.89	8.52	22.88	10.35	31.56	9.35
DualCoOp [15]	0.065	P	51.26	7.93	17.27	8.55	25.83	8.23
MaPLe [8]	3.555	P	58.89	21.85	11.35	18.23	19.03	19.87
CLIP-Adapter [3]	0.131	A	0.32	5.76	19.62	8.00	0.62	6.70
CLIP-LoRA [23]	0.184	A	57.04	18.04	16.13	19.22	25.15	18.61
SurgVLP [20]	0	ZS	2.38	5.81	8.43	7.45	3.71	6.53
HecVLP [21]	0	ZS	0.92	5.82	4.78	7.48	1.55	6.54
PeskaVLP [22]	0	ZS	3.97	6.11	16.75	8.98	6.42	7.28
SurgVLP [20]	133.39	FT	54.75	13.68	17.73	17.46	26.79	15.34
HecVLP [21]	133.39	FT	55.11	12.53	26.55	14.28	35.83	13.35
PeskaVLP [22]	133.39	FT	56.99	16.72	30.35	16.17	39.60	16.44
fine-CLIP (Ours)	3.019	PA	**61.71**	**31.72**	**39.78**	**32.17**	**48.38**	**31.95**

3.4. Results and Discussion

Baselines: We prepare a comprehensive set of baselines, including prompt learning methods like CoOp [24], DualCoOp [15], and MaPLe [8], along with adaptation strategies such as CLIP-Adapter [3] and CLIP-LoRA [23]. We also include SurgVLP [20] and its variants that have been trained on a large corpus of surgical image and text pairs and have shown improved zero-shot phase recognition. We finetune CLIP (CLIP-FT) and SurgVLP variants (denoted by FT) on the images of base triplets for comprehensive evaluation.

Unseen-Target (UT): As shown in Table 1, fine-CLIP with only 3.019M trainable parameters surpasses PeskaVLP (the best SurgVLP variant) by +9.43 percentage points (pp) in novel class F1@3, while showing even larger gains over prompt learning (+16.90 pp vs CoOp) and adaptation strategies (+23.65 pp vs CLIP-LoRA). This superior generalization yields +8.78 pp gain in harmonic mean F1@3 over PeskaVLP.

Unseen-Instrument-Verb (UIV): Table 2 presents results for this challenging setting where models must generalize to unseen instrument-verb combinations. fine-CLIP demonstrates strong generalization to novel combinations, surpassing HecVLP by +2.48 pp in F1@3, while showing substantial gains over prompt learning (+17.74 pp vs MaPLe) and adaptation strategies (+16.10 pp vs CLIP-Adapter). This superior performance on novel combinations is achieved while preserving the best base performance, resulting in a +2.95 pp gain in harmonic mean F1@3 over HecVLP. Across both settings, fine-CLIP demonstrates impressive zero-shot generalization, outperforming both lightweight adaptation methods (CoOp, LoRA) and heavily pretrained models (SurgVLP variants). This performance highlights the value of learning object-centric features and jointly tuning text and vision encoders for generalizing to novel surgical triplets.

Analysis of Zero-shot generalization: As observed from Figure 2, fine-CLIP exhibits distinct generalization patterns

across settings. In UT setting (a), it excels at precise surgical actions on anatomically related structures, with ⟨clipper, clip, cystic-artery⟩ and ⟨scissors, cut, cystic-artery⟩ achieving high performance (~80% and ~70% mAP), benefiting from similar actions on cystic duct during training. Strong generalization is also seen for ⟨irrigator, aspirate, fluid⟩ (~65% mAP), while struggling with targets requiring different interaction patterns in ⟨grasper, grasp, peritoneum⟩ (~5% mAP). In UIV setting (b), the model effectively transfers specialized actions like ⟨scissors, cut, cystic-duct⟩ (~33% mAP) and ⟨irrigator, aspirate, fluid⟩ (~32% mAP), but shows limited performance on retract actions across different instruments (⟨bipolar, retract, liver⟩, ⟨grasper, retract, omentum⟩, ~5-17% mAP). This suggests fine-CLIP's effectiveness depends on both anatomical similarities in UT and action complexity in UIV. Figure 3 visualizes the clusters in the test set of novel triplets, illustrating how SGC module effectively aggregates patches associated with instrument-tissue interactions.

3.5. Ablation Study:

Table 3(a) shows the impact of fine-CLIP components, revealing that incorporating hierarchical loss and object-centric features via SGC enhances zero-shot triplet generalization. Table 3(b) evaluates layer-wise performance with optimal results at the 8th ViT-B/16 layer. Table 3(c) highlights model results across different cluster counts k.

4. Conclusion

In this work, we address the limitations of surgical vision-language models and CLIP's prompt learning and adaptation strategies for zero-shot triplet classification. Existing methods overlook crucial patch-level object semantics by relying on flattened image features. We introduce fine-CLIP, a novel model that overcomes these limitations in three key

Table 2. Zero-shot triplet classification performance (F1@3 and AP in %) on UIV setting. (P: prompt learning, A: adaptation, ZS: zero-shot, FT: Fine-tuning).

| Method	#Trainable Params(MB)	Category	Base		Novel		HM	
			F1@3	mAP	F1@3	mAP	F1@3	mAP
CLIP [12]	0	ZS	0.22	4.88	0.18	6.31	0.2	5.5
CLIP [12]	86.192	FT	39.84	16.33	2.12	7.55	4.03	10.32
CoOp [24]	0.004	P	37.91	11.94	1.94	7.77	3.69	9.42
DualCoOp [15]	0.065	P	37.81	12.41	2.06	6.8	3.91	8.79
MaPLe [8]	3.555	P	44.15	21.19	5.7	8.19	10.09	11.81
CLIP-Adapter [3]	0.131	A	3.16	6.47	7.34	6.98	4.42	6.72
CLIP-LoRA [23]	0.184	A	43.41	20.61	5.05	7.31	9.05	10.79
SurgVLP [20]	0	ZS	2.50	5.65	1.95	6.24	2.19	5.93
HecVLP [21]	0	ZS	2.63	6.12	4.70	6.50	3.37	6.30
PeskaVLP [22]	0	ZS	5.93	6.25	20.73	7.03	9.23	6.62
SurgVLP [20]	133.39	FT	41.21	13.83	13.62	8.05	20.48	10.18
HecVLP [21]	133.39	FT	41.21	13.8	20.96	7.63	27.79	9.83
PeskaVLP [22]	133.39	FT	41.62	17.86	19.94	9.21	26.96	12.15
fine-CLIP (Ours)	3.019	PA	**44.66**	**27.33**	**23.44**	**10.98**	**30.74**	**15.66**
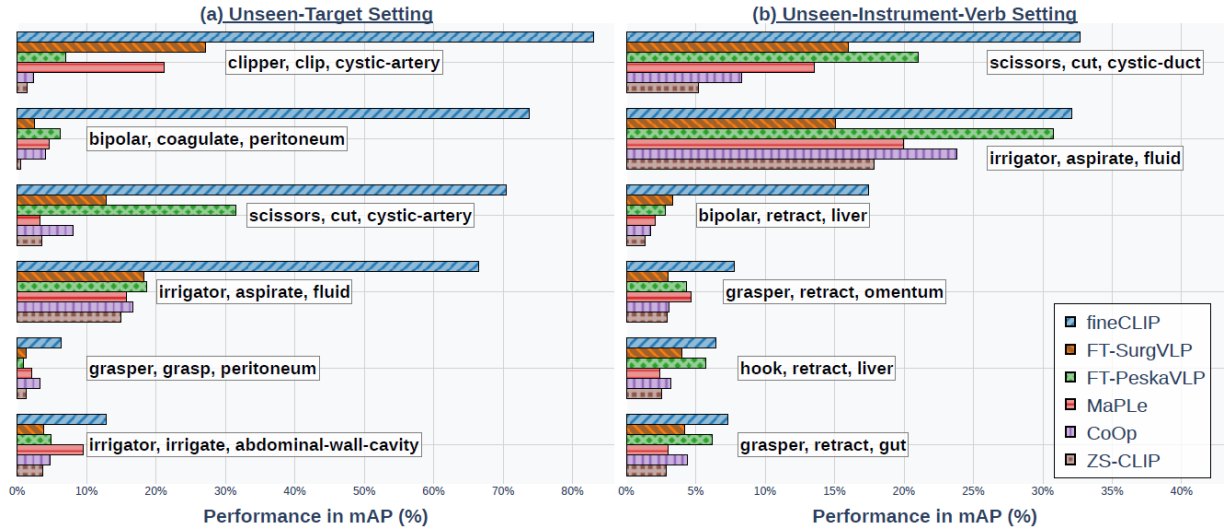
Fig. 2. Performance on novel triplets in (a) UT and (b) UIV settings.

ways: it learns object-centric features from a semantic patch graph using a graph condensation strategy, employs hierarchical prompts to capture hierarchical relationships, and presents a new base-to-novel generalization benchmark for evaluation. This benchmark assesses zero-shot triplet classification performance when trained on a subset of triplet classes, showing improved results over strong baselines.

5. Acknowledgements

Code availability Source code will be provided at `https://github.com/CAMMA-public/fine-CLIP`.

References

[1] Dosovitskiy A, Beyer L, et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR 2021

[2] Gandelsman Y, Efros AA, Steinhardt J (2023) Interpreting clip's image representation via text-based decomposition. arXiv preprint arXiv:231005916

[3] Gao P, Geng S, Zhang R, et al (2024) Clip-adapter: Better vision-language models with feature adapters. IJCV 132(2):581–595

[4] Honarmand M, Jamal MA, Mohareri O (2024) Vidlpro: A video-language pre-training framework for robotic and laparoscopic surgery. In: Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond

[5] Hu EJ, Shen Y, Wallis P, et al (2021) Lora: Low-rank adaptation of large language models. URL `https://arxiv.org/abs/2106.09685`, 2106.09685
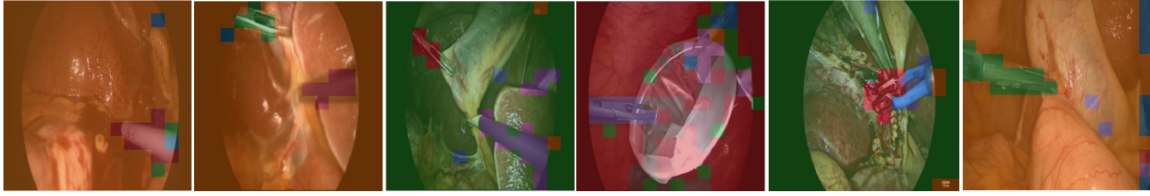
Fig. 3. Qualitative Results: Visualization of the clusters. (Best viewed in color)

Table 3. Ablation Studies on fine-CLIP (Harmonic Mean: F1@3 and AP in %).

3(a): Component-wise Results						3(b): Per-Layer Results			3(c): #Clusters (k)		
SoftPrompt	LoRA	Hierarchy	SGC	F1@3	mAP	j	F1@3	mAP	k	F1@3	mAP
✓				31.56	9.31	7	43.66	27.38	5	47.02	28.50
	✓			25.15	18.61	8	**48.38**	**31.95**	10	43.88	26.08
✓	✓			38.44	20.61	9	33.97	28.20	15	**48.38**	**31.95**
✓	✓	✓		39.16	23.53	10	44.57	28.12	20	41.62	27.62
✓	✓	✓	✓	**48.38**	**31.95**	11	42.17	24.93	25	43.90	28.87

[6] Hu M, Yuan K, Shen Y, et al (2024) Ophclip: Hierarchical retrieval-augmented learning for ophthalmic surgical video-language pretraining. arXiv preprint arXiv:241115421

[7] Jin W, Zhao L, Zhang S, et al (2021) Graph condensation for graph neural networks. arXiv preprint arXiv:211007580

[8] Khattak MU, Rasheed H, Maaz M, et al (2023) Maple: Multi-modal prompt learning. In: CVPR, pp 19113–19122

[9] Lavanchy JL, Ramesh S, Dall'Alba D, et al (2024) Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. IJCARS

[10] Nwoye CI, Padoy N (2022) Data splits and metrics for method benchmarking on surgical action triplet datasets. arXiv preprint arXiv:220405235

[11] Nwoye CI, Yu T, Gonzalez C, et al (2022) Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. MedIA 78:102433

[12] Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: ICML, PMLR, pp 8748–8763

[13] Sharma S, Nwoye CI, Mutter D, et al (2023) Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. IJCARS 18(6):1053–1059

[14] Sharma S, Nwoye CI, Mutter D, et al (2023) Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. MICCAI

[15] Sun X, Hu P, Saenko K (2022) Dualcoop: Fast adaptation to multi-label recognition with limited annotations. NeurIPS 35:30569–30582

[16] Twinanda AP, Shehata S, Mutter D, et al (2016) Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE TMI 36(1):86–97

[17] Veličković P, Cucurull G, Casanova A, et al (2017) Graph attention networks. arXiv preprint arXiv:171010903

[18] Vendrov I, Kiros R, Fidler S, et al (2016) Order-embeddings of images and language. URL https://arxiv.org/abs/1511.06361, 1511.06361

[19] Vercauteren T, Unberath M, Padoy N, et al (2019) Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. Proceedings of the IEEE 108(1):198–214

[20] Yuan K, Srivastav V, Yu T, et al (2023) Learning multi-modal representations by watching hundreds of surgical video lectures. arXiv preprint arXiv:230715220

[21] Yuan K, Srivastav V, Navab N, et al (2024) Hecvl: hierarchical video-language pretraining for zero-shot surgical phase recognition. In: MICCAI

[22] Yuan K, Srivastav V, Navab N, et al (2024) Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. arXiv preprint arXiv:241000263

[23] Zanella M, Ben Ayed I (2024) Low-rank few-shot adaptation of vision-language models. In: CVPR, pp 1593–1603

[24] Zhou K, Yang J, Loy CC, et al (2022) Learning to prompt for vision-language models. IJCV 130(9):2337–2348