

## RESEARCH ARTICLE

# ALF-Surg: Atomic language-guided few-shot surgical phase recognition

Houlong He, Lin Mao, Chengli Song

Shanghai Institute for Minimally Invasive Therapy, School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.

**Corresponding author:** Chengli Song.

**Address correspondence to:** Chengli Song, Shanghai Institute for Minimally Invasive Therapy, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu District, Shanghai 200093, China.  
Tel: +86-021-55272107.  
E-mail: csong@usst.edu.cn.

Received January 23, 2026; Accepted March 25, 2026; Published September 30, 2026

DOI: 10.61189/436470jdqdrx

**Abstract**

**Objective:** To enhance the generalization and transferability of visual-language models (VLMs) pre-trained with extensive clinical data for surgical phase recognition (SPR) and workflow analysis tasks, this paper proposes Atomic Language-Guided Few-Shot Surgical Phase Recognition (ALF-Surg), a lightweight framework designed for efficient adaptation of VLMs to diverse procedures and clinical sites under few-shot settings.

**Methods:** We leverage the rich medical knowledge contained in large language models to transform high-level annotations into vision-based atomic action descriptions, thereby injecting rich spatiotemporal cues and instrument-tissue interaction semantics. A fine-grained multimodal fusion module is applied to jointly encode vision–language features, forming stable and discriminative atomic-level and class-level prototypes. Furthermore, we propose a multi-level multimodal matching strategy, which performs video–video and video–text alignment at both atomic and phase levels to ensure robust decision-making.

**Results:** ALF-Surg was extensively evaluated on three multi-institutional and multi-procedure datasets, including Cholec80, BernBypass70, and StrasBypass70. Compared with zero-shot transfer, ALF-Surg achieves significant accuracy improvements using only eight labeled frames per class (1-shot gains of +13.65%, +30.00%, and +35.03%, respectively). Compared with state-of-the-art few-shot baselines, ALF-Surg consistently establishes new performance records across these datasets. **Conclusion:** Experimental results demonstrate that ALF-Surg achieves robust SPR with minimal supervision across diverse surgical settings.

**Keywords:** Vision–language pre-training, Few-shot learning, Surgical video analysis

**Highlights**

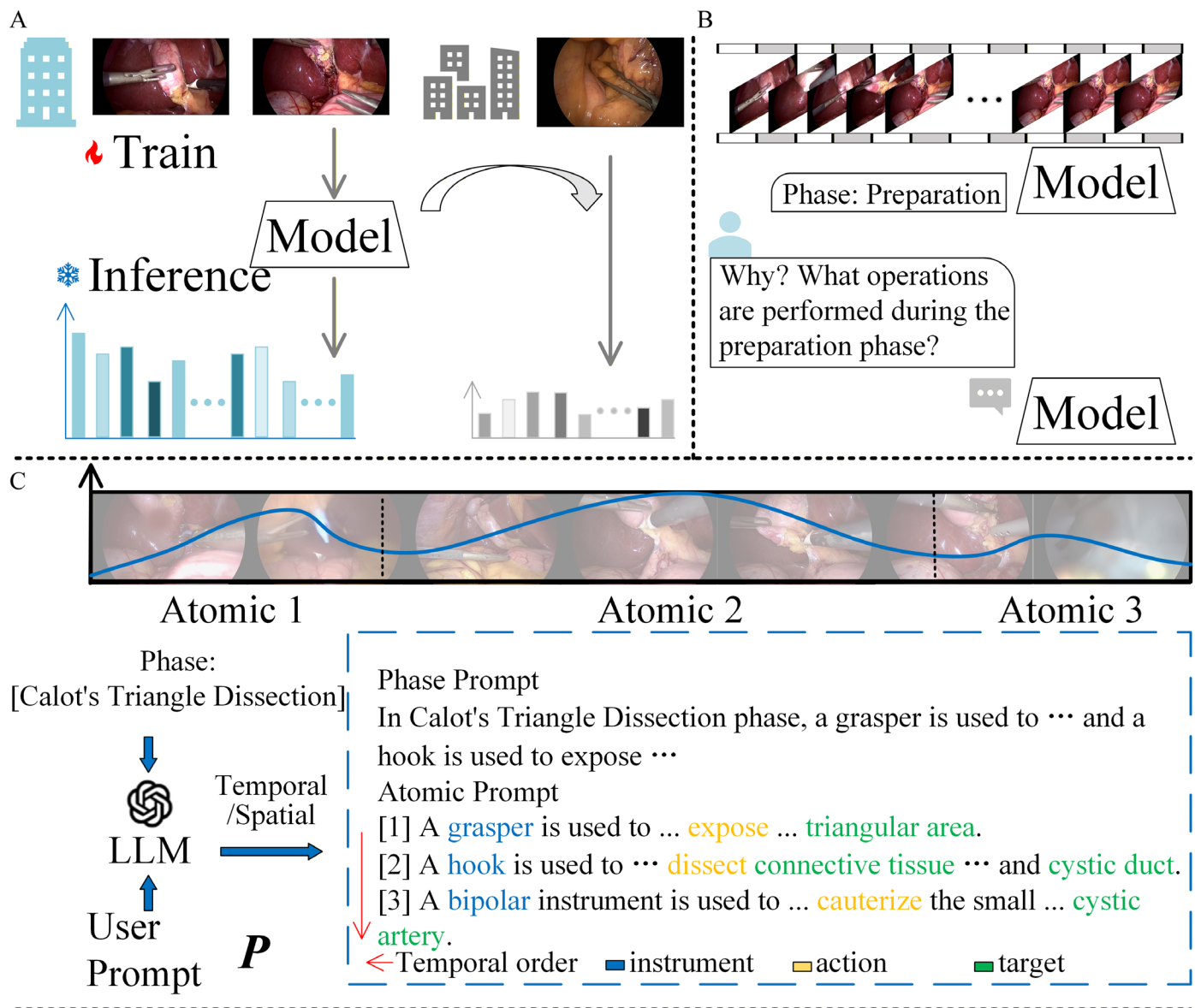
- Leverages large language model-driven atomic descriptions to inject fine-grained temporal and interaction semantics into phase labels.
- Proposes a multi-level multimodal matching strategy to align video-text representations at both atomic and class levels.
- Enables efficient few-shot adaptation to diverse surgical procedures and clinical sites with minimal annotation costs.
- Establishes new state-of-the-art performance in few-shot learning on three multi-institutional datasets, demonstrating robust cross-domain generalization.

**1 INTRODUCTION**

The rapid development of surgical computer vision has greatly facilitated the emergence of intelligent operating rooms and

artificial intelligence-assisted surgical systems [1, 2]. Recent research has evolved from task-specific models tailored to single surgical procedures or institutions toward more generalizable frameworks capable of handling diverse surgical scenes





**Figure 1. Motivation for atomic language-guided surgical phase recognition.** (A) Traditional single-center model training and transfer; (B) Traditional class labels lack semantic depth; (C) An LLM analyzes phase labels to generate atomic-level descriptions. LLM, large language model.

and tasks [3-5]. Such generalization capability is crucial for the practical deployment of intelligent surgical systems across different hospitals and surgical procedures.

Surgical phase recognition (SPR) serves as a fundamental component in surgical workflow analysis, enabling real-time assessment of surgical progress, skill evaluation, and intraoperative decision support [6, 7]. Despite promising advances, current SPR models still face two major limitations that hinder clinical deployment. First, existing methods typically rely on single-center datasets and task-specific architectures. Due to variations in surgical styles, patient anatomies, and recording conditions across hospitals, these models

often fail to generalize to unseen domains, as shown in **Figure 1A**.

Second, most current SPR models rely on discrete phase labels for supervision. These labels encode limited semantic information, whereas each surgical phase involves highly complex procedures. For example, in cholecystectomy, the “Calot’s Triangle Dissection” phase requires first using forceps to retract the gallbladder and expose Calot’s triangle, then separating the connective tissue around the gallbladder plate and cystic duct, clarifying the anatomical relationship between the cystic duct and cystic artery, avoiding collateral damage, and preparing for subsequent clamping, as shown in **Figure 1B, 1C**.

Few-shot learning is an effective solution to the domain transfer problem in surgical tasks [8, 9]. By fine-tuning large-scale visual-language pre-trained models (VLPMs), such as Contrastive Language–Image Pre-training (CLIP), Surgical Vision–Language Pre-training (SurgVLP), and Ophthalmic Surgical Video–Language Pre-training, on a small number of labeled samples, rapid model adaptation can be achieved while reducing annotation costs [10–12]. Based on advances in few-shot image classification, many existing methods employ a metric-based meta-learning paradigm by learning a transferable embedding space and a similarity function to measure the distance between query and support samples [13, 14]. The key is how to learn discriminative and transferable representations in complex surgical workflows.

To bridge the substantial domain gap between natural and surgical scenarios, recent efforts have focused on developing domain-specific visual-language models (VLMs). In the surgical field, SurgVLP represents a milestone toward foundation models for surgery [11]. It constructs a large-scale multimodal dataset (Surgical Video Lecture Pre-training dataset) from publicly available surgical video lectures and generates synchronized textual transcripts via automatic speech recognition systems. SurgVLP also introduces a multi-view contrastive objective that aligns video clips with corresponding textual descriptions, enabling the learning of temporally coherent, cross-modal embeddings. Similarly, Ophthalmic Surgical Video–Language Pre-training, which targets ophthalmic surgery, utilizes surgical videos from YouTube and automatic speech recognition-based transcribed texts and designs multiple training objectives to capture hierarchical and fine-grained representations [12]. These domain-specific VLMs have demonstrated zero-shot and few-shot transfer capabilities in various surgical visual-language tasks, including text-based video retrieval, temporal activity localization, video caption generation, and instrument recognition.

To enhance the representational capabilities of few-shot models, researchers are increasingly incorporating multimodal cues as complementary information [15–19]. For example, Tip-Adapter utilizes cached image and text embedding key-value pairs to construct a non-parametric dual-cache model, achieving zero-shot or few-shot transfer without additional backpropagation [20]. Similarly, in medical image processing, Linear Probing (LP)+Text fuses textual priors with visual prototypes to learn an enhanced prototype composed of both visual and textual information [21]. While these methods demonstrate significant potential, they are primarily designed for static image classification or short video clips and are not suitable for tasks such as SPR, which involve complex visual dynamics and sophisticated tool-tissue interactions.

To overcome these limitations, this paper proposes a novel framework, Atomic Language-Guided Few-Shot Surgical Phase Recognition (ALF-Surg). ALF-Surg is a lightweight

adaptive framework. As shown in **Figure 2**, this framework performs few-shot transfer based on a frozen VLPM, enabling the base model to generalize across procedures and institutions with minimal supervision and training costs. Second, to go beyond traditional label semantics, we leverage a large language model (LLM) to introduce an atomic language-guided supervision mechanism. Surgical phases consist of a series of atomic actions that collectively define the surgical intent. Therefore, each phase can be decomposed into atomic descriptions (e.g., “expose the cystic duct” and “dissect Calot’s triangle”), providing a richer and more structured linguistic foundation, as shown in **Figure 1C**. Simultaneously, a fine-grained multimodal fusion module integrates textual and visual features at the atomic level to construct modality-aligned category prototypes. To ensure robust few-shot matching, we also propose a multi-level multimodal matching module that performs complementary video-video and video-text matching at both the atomic and class levels. Validated using three datasets from different institutions—Cholec80, BernBypass70, and StrasBypass70—containing different surgical procedures (laparoscopic cholecystectomy and gastric bypass surgery), ALF-Surg achieves state-of-the-art (SOTA) performance by fully decomposing and aligning surgical phases, significantly improving the model’s generalization and transfer capabilities in the field of SPR [22, 23].

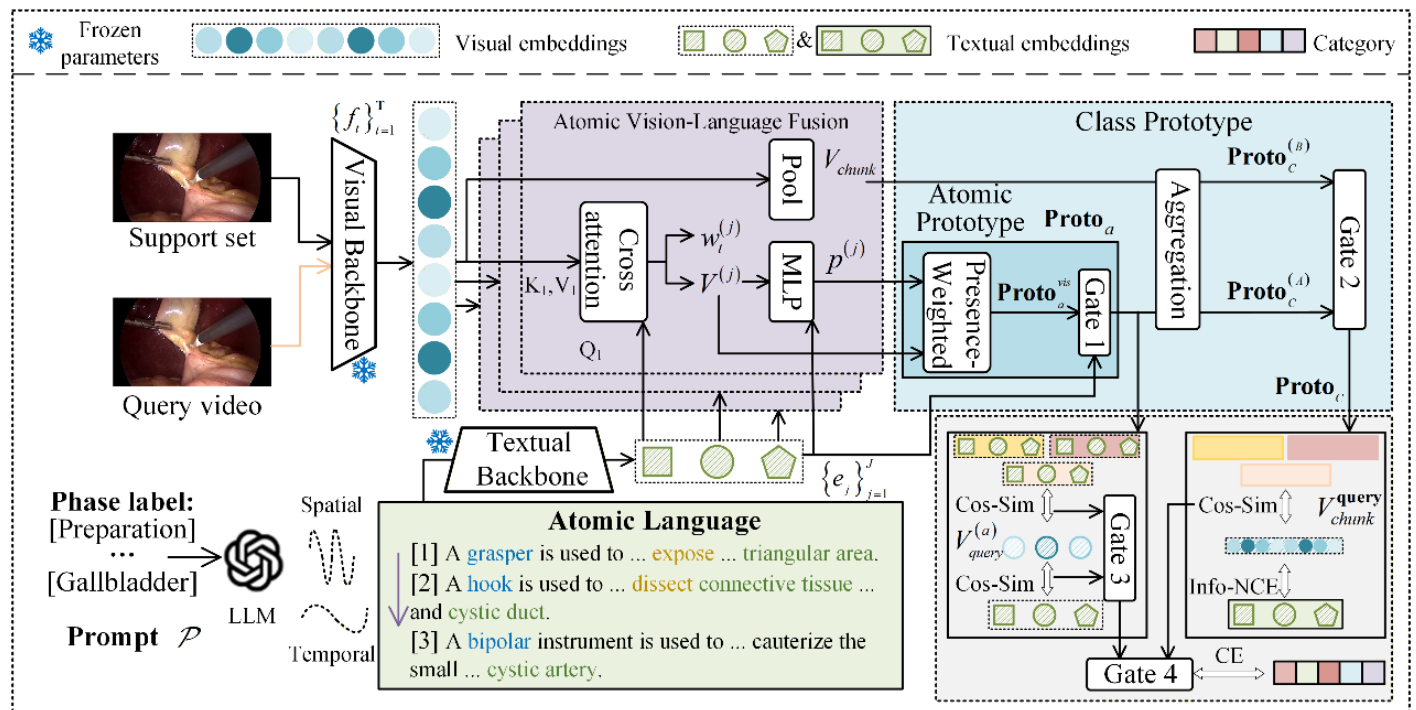
In summary, this study makes three main contributions. First, we propose ALF-Surg, a novel, lightweight, and adaptive framework for SPR that is designed to transfer visual-language models to new surgical domains with minimal supervision. Second, the proposed fine-grained multimodal fusion module and multi-level multimodal matching module jointly encode atomic-level text and visual embeddings to construct semantically aligned prototypes and perform complementary video-video and video-text matching at the atomic and class levels. Third, validation results on three datasets involving different institutions and procedures, including Cholec80, BernBypass70, and StrasBypass70, demonstrate that ALF-Surg achieves SOTA performance in both model transfer and generalization.

## 2 MATERIALS AND METHODS

### 2.1 Overview

#### 2.1.1 Problem definition

The goal of few-shot SPR transfer is to adapt a pre-trained vision–language model to new surgical procedures or institutions using only a limited number of labeled samples. Formally, the target institution provides a small set of labeled examples covering  $N$  surgical phases, which is referred to as the support set. Each sample corresponds to a short video chunk (e.g., four or eight consecutive frames) that captures coherent temporal and contextual cues within a phase, and each phase contains  $K$



**Figure 2. Overview of the proposed ALF-Surg framework.** It mainly consists of surgical phase text anatomy, a fine-grained multimodal fusion module, and a multi-level multimodal matching module. ALF-Surg, Atomic Language-Guided Few-Shot Surgical Phase Recognition; LLM, large language model; MLP, multilayer perceptron; CE, cross-entropy; Info-NCE, information noise-contrastive estimation; Cos-Sim, cosine similarity.

such chunks. During inference, the model classifies unlabeled query videos by matching them to the prototypes derived from the support set, without extensive retraining or annotation.

### 2.1.2 Overall framework

As shown in **Figure 2**, the proposed ALF-Surg framework consists of three main parts: surgical phase text anatomy, a fine-grained multimodal fusion module, and a multi-level multimodal matching module. First, we leverage the rich prior knowledge of LLMs to decompose coarse surgical phase labels into a series of atomic-level textual descriptions. These textual descriptions are encoded by a frozen text backbone to obtain atomic text embeddings. Simultaneously, the frozen visual backbone extracts frame-wise visual features from the surgical video. Next, in the fine-grained multimodal fusion module, visual and textual embeddings are jointly fused to construct multi-level prototypes that encode both atomic and class-level semantics. Finally, in the multi-level multimodal matching module, query video representations are compared against support prototypes at both the atomic and class levels, performing complementary video-video and video-text matching to ensure robust recognition under few-shot conditions. In addition, four adaptive gating mechanisms are introduced throughout the framework to dynamically regulate the contribution of each module.

### 2.2 Surgical phase text anatomy

With the advancement of LLMs, it has become possible to leverage their rich domain knowledge to automatically decompose phase labels into a sequence of atomic, semantically explicit surgical actions. To achieve this, we designed a set of task-specific prompts that guide the LLM to expand each surgical phase label into a sequence of clinically valid atomic steps strictly aligned with standard surgical guidelines. Finally, each atomic description is encoded through the frozen text encoder, producing a set of textual embeddings (**Figure 2**). These generated descriptions and their pre-computed embeddings remain fixed during training to ensure the reproducibility of our framework:

$$E_a = \{e_1, e_2, \dots, e_j\}, e_j \in \mathbb{R}^D \quad (1)$$

where  $j$  denotes the number of atomic textual descriptions for a given phase, and  $D$  denotes the embedding dimension.

### 2.3 Fine-grained multimodal fusion module

The fine-grained multimodal fusion module performs atomic vision-language fusion. Given the atomic-level text descriptions  $E_a$  and the frame-wise visual features  $F_r = \{f_1, f_2, \dots, f_j\}$ , the goal of the atomic vision-language fusion module is to align

the fine-grained semantics between surgical actions described by atomic language and visual patterns observed in video frames and to construct robust atomic-level multimodal representations for subsequent prototype generation and matching. As illustrated in **Figure 2**, we adopt a cross-attention-based fusion strategy to achieve this correspondence. Specifically, each atomic textual embedding  $e_j$  is treated as a query that attends to all frame-wise visual features  $f_i$ , which serve as the key and value. Through this operation, we obtain a set of attention weights  $w_t^j$  that quantify the temporal relevance of each frame to the atomic query and a multimodal fused representation  $V^{(j)}$  that integrates visual information guided by the atomic textual semantics:

$$\alpha_t^j = \frac{\langle W_q e_j, W_k f_t \rangle}{\sqrt{d}} \quad w_t^j = \frac{\exp(\alpha_t^j / \tau)}{\sum_{u=1}^T \exp(\alpha_u^j / \tau)}, \quad V^{(j)} = \sum_{t=1}^T w_t^j W_v f_t \quad (2)$$

where  $W$  is a learnable linear layer,  $\sqrt{d}$  is the scaling factor, and  $\tau$  is the temperature parameter.

To further assess the reliability of each atomic-visual pair, we concatenate  $V^{(j)}$  and  $e_j$  and feed the result into a lightweight multilayer perceptron to obtain a presence score  $p^{(j)} \in [0, 1]$ . This score measures the confidence that the corresponding atomic action is visually manifested in the given clip. In subsequent modules,  $p^{(j)}$  serves as a soft reliability indicator that adaptively modulates the contribution of each atomic prototype during prototype fusion and final classification.

In parallel, we perform temporal pooling over  $F_t$  to obtain a chunk-level visual representation  $V_{chunk}$ , which captures the overall appearance and contextual background of the surgical phase.

## 2.4 Atomic-level prototype construction

We construct atomic-level prototypes based on  $p^{(j)}$ ,  $V^{(j)}$ , and  $E_a$ . These prototypes aggregate the semantic and visual evidence of the same atomic concept from support data. We adopt a presence-weighted aggregation strategy that accounts for the reliability of each support instance. Specifically, for a given atomic operation  $a$  and its  $K$  support samples in each phase, we first compute the visual prototype  $\text{Proto}_a^{\text{vis}}$ , which is a weighted average of the multimodal fusion representations, with the presence score serving as the soft confidence weight:

$$\text{Proto}_a^{\text{vis}} = \frac{\sum_{k=1}^K p_k^{(a)} V_k^{(a)}}{\sum_{k=1}^K p_k^{(a)} + \varepsilon} \quad (3)$$

This aggregation ensures that more reliable instances contribute more strongly to the prototype.

However, in the early stages of model training, visual evidence may be unreliable due to poor visual-text alignment, resulting in low presence scores. To mitigate this cold-start issue, we introduce a textual fallback mechanism that adaptively integrates the atomic textual embedding  $E_a$  into the prototype representation. The final atomic-level prototype is defined as follows:

$$\bar{p}^{(a)} = \frac{1}{K} \sum_k p_k^{(a)}, \quad \text{gate1} = \sigma(\alpha \cdot \bar{p}^{(a)} + b), \quad (4)$$

$$\text{Proto}_a = \text{gate1} \cdot \text{Proto}_a^{\text{vis}} + (1 - \text{gate1}) \cdot E_a$$

where  $\text{gate1}$  is a sigmoid-based adaptive coefficient controlled by the average presence score over the  $K$  support samples, and  $\alpha$  and  $b$  are learnable parameters, as shown in **Figure 3A**.

As training progresses and visual alignment improves,  $\text{gate1}$  gradually approaches 1, allowing the prototype to shift toward the visual representation  $\text{Proto}_a^{\text{vis}}$ . The resulting atomic prototype combines the advantages of both modalities: text embeddings provide a semantic foundation and robustness to domain variations, while visual features ensure sensitivity to surgical appearance and procedural variations.

## 2.5 Class-level prototype construction

The class-level prototype serves as the final category reference during inference. A reliable prototype should simultaneously provide semantic interpretability, visual discriminability, and robustness against noise or missing evidence. To achieve this goal, we construct class-level prototypes using a dual-path fusion scheme with an adaptive gating mechanism.

### 2.5.1 Semantic-driven path

Given the set of atomic-level prototypes  $\text{Proto}_a$  belonging to surgical phase  $C$ , we aggregate them into a semantic class prototype:

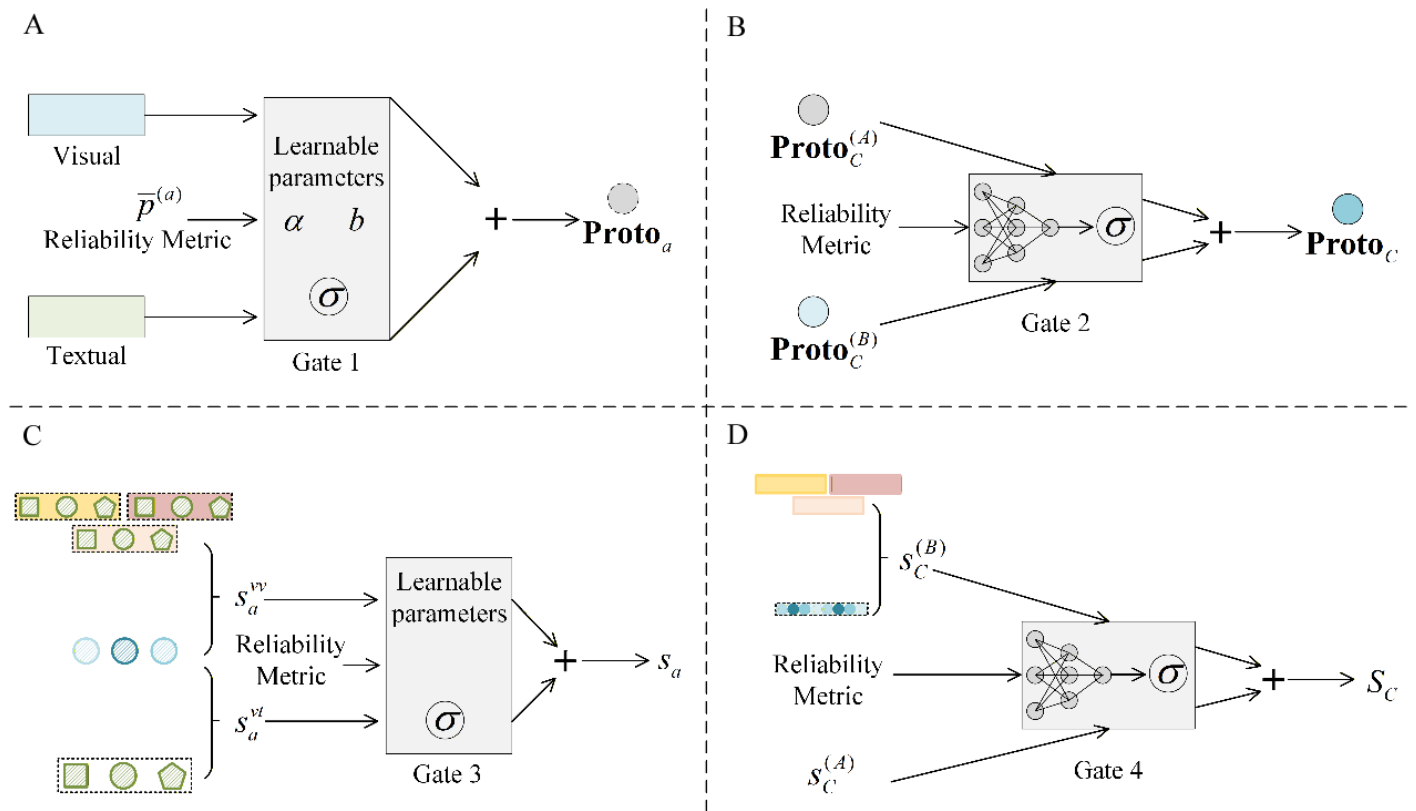
$$\text{Proto}_C^{(A)} = \frac{1}{|A_C|} \sum_{a \in A_C} \text{Proto}_a \quad (5)$$

where  $A_C$  denotes the atomic concepts corresponding to phase  $C$ . This aggregation preserves the fine-grained semantics introduced by LLM-derived atomic descriptions.

### 2.5.2 Visual-driven path

In parallel, we construct a visual class prototype by averaging the chunk-level visual embeddings  $V_{chunk}$  from the  $K$ -shot support set:

$$\text{Proto}_C^{(B)} = \frac{1}{K} \sum_{k=1}^K V_{chunk,k} \quad (6)$$



**Figure 3. Adaptive gating mechanisms.** (A) Gate 1: Atomic-level prototype construction; (B) Gate 2: Class-level prototype construction; (C) Gate 3: Atomic-level score fusion; (D) Gate 4: Class-level score fusion. Proto, prototype.

This path captures the global visual environment of the surgical phase, providing strong discriminative cues and supplementing atomic semantics.

### 2.5.3 Adaptive fusion via reliability-aware gating

Using only atomic aggregation may reduce discriminative power when visual evidence is weak or atomic descriptions are insufficient, while using only visual aggregation may weaken semantic alignment capabilities across domains. To address this, we design a learnable adaptive gating system to dynamically fuse the two types of information, as shown in **Figure 3B**:

$$\text{gate2} = \sigma(\text{MLP}([\bar{p}_C, \text{coverage}_C, \text{var}(p_C)]))$$

$$\text{Proto}_C = \text{gate2} \cdot \text{Proto}_C^{(A)} + (1 - \text{gate2}) \cdot \text{Proto}_C^{(B)} \quad (7)$$

where gate2 represents the gating mechanism, and  $\bar{p}_C$ ,  $\text{coverage}_C$ , and  $\text{var}(p_C)$  represent the three gating inputs.  $\bar{p}_C$  represents the average presence score of all atomic concepts in phase C across the  $K$ -shot support samples, reflecting the overall visual support for semantic evidence. A high value indicates that the atomic actions have strong visual evidence and are more likely to resemble visual prototypes.  $\text{coverage}_C$  represents the proportion of atomic concepts whose presence scores exceed the threshold  $\tau=0.5$ . If only a small number of atomic

concepts are detected (low  $\text{coverage}_C$ ), more reliance should be placed on global visual cues.  $\text{var}(p_C)$  represents the variance. High variance indicates inconsistent evidence, and caution should be exercised when fusing the two types of information. For instance, in challenging surgical scenarios in which the field of view is obscured by smoke, bleeding, or lens blurring, the visual evidence becomes unstable (indicated by high  $\text{var}(p_C)$  or low  $\bar{p}_C$ ). In such cases, gate2 automatically prioritizes the stable, domain-agnostic textual atomic prototypes over unreliable visual features to maintain robust recognition. The multilayer perceptron learns the nonlinear interactions among these factors, enabling gate2 to dynamically adjust according to different domains, surgical styles, and support set quality to construct reliable class-level prototypes  $\text{Proto}_C$ .

### 2.6 Multi-level multimodal matching

In few-shot learning scenarios, traditional methods typically rely on distance metrics between visual features for matching, neglecting the high-value semantic information contained in textual descriptions. This is especially true in surgical videos, where visual features in different surgical phases may be similar, but their semantic behaviors can be drastically different. To fully utilize visual-linguistic cues, ALF-Surg employs video-video and video-text matching at both the atomic and class levels.

### 2.6.1 Atomic-level matching

At the atomic level, ALF-Surg matches the query video with all atomic concepts associated with that category. The model computes similarity in both video-video and video-text matching modalities and then fuses them using a self-adjusting gate (gate3) to obtain atomic-level logits, as shown in **Figure 3C**.

Given a query sample, we first obtain its fused visual representation  $V_{query}^{(a)}$  through atomic vision–language fusion. Then, we calculate its similarity to each atomic-level prototype. Similarly, we measure the alignment between the query video and the atomic textual embedding using cosine similarity:

$$s_a^{vv} = \cos(V_{query}^{(a)}, \text{Proto}_a), \quad s_a^{vt} = \cos(V_{query}^{(a)}, E_a) \quad (8)$$

where  $\cos$  denotes the cosine similarity function. For simplicity, the normalization operator has been omitted. Scores  $s_a^{vv}$  and  $s_a^{vt}$  reflect whether the query video visually contains fine-grained operation patterns represented by atomic-level prototypes, enabling the model to supplement the visual scene with atomic-level descriptions. These two scores are then fused using a self-adjusting gating system (gate3) based on the support set:

$$\begin{aligned} \text{support\_conf}_a &= \text{clip}(\bar{p}_a^\lambda \cdot \bar{z}_a^{(1-\lambda)/2} \cdot \bar{c}_a^{(1-\lambda)/2}, 0, 1) \\ \text{gate3} &= \sigma(\alpha \cdot \text{support\_conf}_a + b) \\ s_a &= \text{gate3} \cdot s_a^{vv} + (1 - \text{gate3}) \cdot s_a^{vt} \\ s_C^{(A)} &= \frac{1}{|A_C|} \sum_{a \in A_C} s_a \end{aligned} \quad (9)$$

gate3 is primarily determined by three factors.  $\bar{p}_a$  represents presence strength; a higher average presence score for the atomic element in the support set indicates more credible visual evidence.  $\bar{z}_a = \frac{1}{K} \sum_k \max_t w_{k,t}^{(a)}$  represents attention concentration, where  $w_{k,t}^{(a)}$  represents the attention weight at frame  $t$  when processing the  $k$ -th sample. If the maximum attention value is concentrated on stable frames, the model consistently finds the key operation. Finally,  $\bar{c}_a = \frac{1}{K} \sum_k \cos(V_k^{(a)}, \text{Proto}_a^{vis})$  represents prototype consistency, measuring the degree of consistency between the visual representations of the same atomic element across the  $K$  support samples and its atomic-visual prototype. These three metrics are combined into  $\text{support\_conf}_a$  through a weighted product and mapped to  $[0, 1]$ .  $\lambda$  is used to control whether  $\bar{p}_a$  or  $\bar{c}_a$  is weighted more heavily. This self-regulating gating design measures the reliability of the atomic-level prototype and reduces the adverse effects of noisy or inconsistent atomic cues on the final category prediction.

### 2.6.2 Class-level matching

At the class level, the model integrates two complementary sources of information to determine the final classification:

aggregated atomic logits  $s_C^{(A)}$  and class-level video-video similarity scores.

Specifically, the video-video matching score is calculated as the cosine similarity between the chunk-level visual representation of each query video and the class-level prototype:

$$s_C^{(B)} = \cos(V_{chunk}^{query}, \text{Proto}_C) \quad (10)$$

This similarity captures the global phase context, including scene layout, dominant surgical tools, and motion patterns. Subsequently, we obtain the final category logits by fusing  $s_C^{(A)}$  with  $s_C^{(B)}$ :

$$\begin{aligned} \text{gate4} &= \sigma(\text{MLP}([\bar{p}_q, H(s_a), \text{support\_conf}_{\text{mean}}])) \\ S_C &= \text{gate4} \cdot s_C^{(B)} + (1 - \text{gate4}) \cdot s_C^{(A)} \end{aligned} \quad (11)$$

The design of gate4 also incorporates three complementary metrics to control whether to place more trust in atomic-level results or to rely more on global visual features during fusion, as shown in **Figure 3D**.  $\bar{p}_q$  represents the mean presence score of the query sample for each atomic concept. Additionally, we calculate the entropy of the atomic matching scores of the query sample  $H(s_a)$ . Low entropy indicates that the model is confident in the results of some atomic operations and that atomic-level results can provide clear discriminative signals.  $\text{support\_conf}_{\text{mean}}$  represents the reliability of the atomic prototype in the support set, obtained by averaging  $\text{support\_conf}_a$ . The fused class-level logits are converted into posterior probabilities via a temperature-scaled softmax:

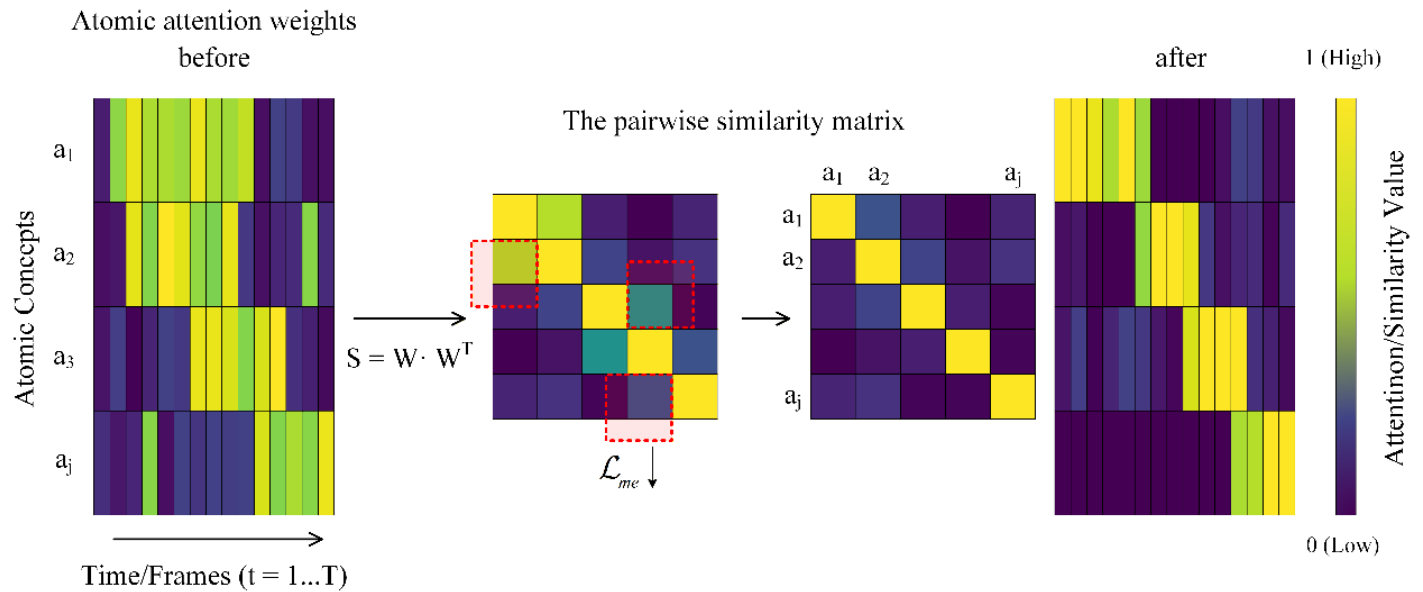
$$P(C | \text{query}) = \text{Softmax}(S_C / \tau) \quad (12)$$

### 2.6.3 Video-text matching

To further enhance the model's visual-text alignment and prototype construction capabilities, we introduce class-level video-text matching via an information noise-contrastive estimation (Info-NCE) contrastive objective during training. Given a query visual representation and the set of class-level textual embeddings  $\{T^C\}_{C=1}^N$ , we compute the similarity score and then calculate the Info-NCE loss, as shown in **Figure 2**:

$$s_C^{vt} = \cos(V_{chunk}^{query}, T^C), \quad L_{vt} = -\log \frac{\exp(s_y^{vt} / \tau_{vt})}{\sum_{C=1}^N \exp(s_C^{vt} / \tau_{vt})} \quad (13)$$

where  $y$  denotes the ground-truth surgical phase of the query sample, and  $\tau_{vt}$  is a learnable temperature. By bringing positive sample pairs  $(V_{chunk}^{query}, T^y)$  as close as possible, this objective helps distinguish the visual representation of the query from erroneous category descriptions.



**Figure 4. Visualization of the temporal decoupling mechanism via the mutual exclusion loss.** Minimizing the similarity values of high-value off-diagonal elements in the similarity matrix, indicated by red dashed boxes, reduces semantic conflicts and enforces a temporally orthogonal distribution of atomic attention.

## 2.7 Loss function

To optimize the proposed ALF-Surg, we supervise it using three objectives: classification accuracy, multimodal alignment, and temporal decoupling. After obtaining the prediction results  $P(C | \text{query})$ , we apply a standard cross-entropy loss:

$$L_{vv} = (\hat{y}, P(C | \text{query})) \quad (14)$$

where  $\hat{y}$  is the surgical phase label. The Info-NCE loss used for class-level video-text matching is denoted as  $L_{vt}$ .

The atomic descriptions generated by the LLM typically do not have an explicit supervisory relationship with the frames contained in each phase. Ideally, different atomic concepts should focus on different points in time within the video. To facilitate this temporal decoupling, we introduce a mutual exclusion loss on the attention maps generated by the atomic vision-language fusion module, as shown in **Figure 4**. Given the attention distributions  $\{w_t^{(1)}, w_t^{(2)}, \dots, w_t^{(j)}\}$  for the  $J$  atomic concepts, we compute the pairwise cosine similarity:

$$L_{me} = \frac{1}{J(J-1)} \sum_{i \neq j} \cos(w^{(i)}, w^{(j)}) \quad (15)$$

Minimizing this loss penalizes excessive overlap between atomic attention maps, thus compensating for the lack of explicit frame-level supervision in atomic descriptions.

The final loss is a weighted combination of the above terms, where the balancing weights are determined empirically via grid search on the validation set ( $\lambda_{vv}=1.0$ ,  $\lambda_{vt}=0.3$ ,  $\lambda_{me}=1.0$ ):

$$L = \lambda_{vv} L_{vv} + \lambda_{vt} L_{vt} + \lambda_{me} L_{me} \quad (16)$$

## 2.8 Datasets and metrics

We evaluate ALF-Surg on three widely used laparoscopic surgery datasets: Cholec80, BernBypass70, and StrasBypass70. Cholec80 contains 80 videos of laparoscopic cholecystectomy. Each video is annotated by expert surgeons with seven predefined surgical phases. BernBypass70 and StrasBypass70 are laparoscopic Roux-en-Y gastric bypass surgery datasets. BernBypass70 includes 70 videos (average duration:  $73 \pm 20$  min) with 8 annotated surgical phases, collected at the University Hospital of Bern, whereas StrasBypass70 contains 70 videos (average duration:  $111 \pm 33$  min) with 10 annotated surgical phases, captured at the University Hospital of Strasbourg. Following common practice in prior studies, we adopt the official data splits for all datasets to ensure fair comparison and reproducibility [22, 23].

## 2.9 Training and inference

We build our framework upon SurgVLP, a publicly available vision-language foundation model pre-trained on large-scale surgical video lectures. To preserve the pre-trained surgical prior and reduce computational overhead, the visual encoder and text encoder are kept frozen during both training and inference. During training, a chunk of eight consecutive frames is treated as one support sample. For each episode, we randomly sample  $K$ -shot chunks for each of the  $N$  surgical phases. During inference, the model performs dense frame-wise prediction on the complete test videos. ALF-Surg is trained end-to-end on a single NVIDIA RTX 5090 graphics processing

**Table 1. Comparison of different methods using accuracy and F1 score (Acc%/F1%) on the three datasets**

| Method                    | Shot setting | Cholec80 Acc%/F1%  | StrasBypass70 Acc%/F1% | BernBypass70 Acc%/F1% |
|---------------------------|--------------|--------------------|------------------------|-----------------------|
| CLIP                      | 0            | 30.80/13.10        | 16.90/5.50             | 14.80/4.10            |
| SurgVLP                   | 0            | 34.70/24.40        | 10.80/6.90             | 11.40/7.20            |
| RN50                      | Full         | 91.20/81.00        | 82.50/71.10            | 74.50/42.40           |
| Linear-Probe (MICCAI2024) | 1            | 35.48/21.15        | <u>32.15/27.96</u>     | 24.33/19.48           |
|                           | 4            | 40.43/29.49        | 42.21/37.45            | 32.35/27.85           |
|                           | 8            | 44.02/32.66        | 46.40/40.52            | 37.62/29.84           |
| LP+Text (MICCAI2024)      | 1            | 33.60/22.26        | 30.83/26.39            | 23.63/18.52           |
|                           | 4            | 40.16/28.49        | 37.69/33.01            | 30.01/26.97           |
|                           | 8            | 42.51/32.36        | 42.61/37.79            | 37.71/29.10           |
| CLIP-Adapter (IJCV23)     | 1            | 31.79/22.42        | 23.73/16.02            | 25.48/22.71           |
|                           | 4            | 38.06/28.42        | 39.43/31.89            | 29.43/26.78           |
|                           | 8            | 38.89/30.03        | 44.26/38.32            | 37.12/30.62           |
| Tip-Adapter-F (ECCV2022)  | 1            | 35.32/24.97        | 28.82/24.01            | 22.25/18.67           |
|                           | 4            | 40.84/30.25        | 36.09/26.95            | 29.25/24.75           |
|                           | 8            | 46.43/37.64        | 37.88/31.40            | 34.44/28.76           |
| Proto-CLIP (IROS2024)     | 1            | 31.48/20.44        | 29.43/26.85            | <u>27.88/24.93</u>    |
|                           | 4            | 38.34/23.07        | 31.86/23.62            | 28.07/28.89           |
|                           | 8            | 40.92/31.63        | 30.13/25.18            | 30.16/29.46           |
| 2SFS (CVPR2025)           | 1            | <u>35.91/32.21</u> | 31.27/26.40            | 27.44/19.65           |
|                           | 4            | <u>49.34/40.34</u> | <u>45.80/37.96</u>     | <u>41.42/27.78</u>    |
|                           | 8            | <u>50.30/43.42</u> | <u>49.37/41.26</u>     | <u>41.35/28.78</u>    |
| ALF-Surg                  | 1            | <b>48.35/39.36</b> | <b>45.83/36.58</b>     | <b>41.40/26.48</b>    |
|                           | 4            | <b>49.31/40.48</b> | <b>49.16/39.40</b>     | <b>42.90/27.73</b>    |
|                           | 8            | <b>50.59/41.37</b> | <b>50.73/40.43</b>     | <b>43.18/28.21</b>    |

Note: For each  $K$ -shot setting, results are averaged over three support sets. Bold values indicate the best results, and underlined values indicate the second-best results. Acc, accuracy; F1, F1 score; CLIP, Contrastive Language–Image Pre-training; SurgVLP, Surgical Vision–Language Pre-training; RN50, ResNet-50; LP, linear probe; Proto-CLIP, a prototypical-network-based method built on CLIP; MICCAI, Medical Image Computing and Computer Assisted Intervention; IJCV, International Journal of Computer Vision; ECCV, European Conference on Computer Vision; IROS, IEEE/RSJ International Conference on Intelligent Robots and Systems; CVPR, Conference on Computer Vision and Pattern Recognition; 2SFS, Two-Stage Few-Shot Adaptation.

unit using the AdamW optimizer with an initial learning rate of  $1 \times 10^{-4}$ . Consistent with prior literature, we report accuracy and F1 score computed on frame-level predictions. All results are averaged over three random support sets for robustness.

### 3 RESULTS

#### 3.1 Comparison with SOTA models

We compare ALF-Surg with three representative types of methods from different perspectives: zero-shot transfer using VLPs, fully supervised models, and few-shot adaptation methods. For a fair comparison, the number of support samples is kept consistent across all few-shot settings. Specifically, 1-shot, 4-shot, and 8-shot correspond to 8, 32, and 64 frames, respectively, matching our chunk-based sampling strategy.

According to the results in **Table 1**, although CLIP benefits from massive image-text pre-training, its direct transfer to sur-

gical domains yields limited performance due to substantial domain and semantic gaps. SurgVLP introduces surgery-specific prior information, but its zero-shot performance is still insufficient for surgical phase inference. ResNet-50 represents the SOTA model trained in a fully supervised manner on each dataset, providing a strong upper bound.

Linear Probe trains a linear classifier using image features [21]. LP+Text enhances LP through text embeddings, slightly improving generalization by introducing semantic priors [21]. CLIP-adapter cleverly balances pre-trained general knowledge and downstream task-specific knowledge by adding a light-weight residual feature adapter [15]. The core idea of Tip-Adapter is to build a feature cache from support samples and efficiently fuse support-set knowledge with the pre-trained CLIP model through feature retrieval [20]. Unlike our method, a prototypical-network-based method built on CLIP constructs image prototypes and text prototypes separately [24]. During classification decisions, query samples are used to simultane-

**Table 2. Ablation study of model modules on the Cholec80 and BernBypass70 datasets**

| SurgVLP      | 0-Shot      |             | Cholec80: 34.70/24.40 |             | BernBypass70: 11.40/7.20 |  |
|--------------|-------------|-------------|-----------------------|-------------|--------------------------|--|
| Atomic-Proto | Matching    |             |                       |             | $L_{me}$                 |  |
|              | Class-level |             | Atomic-level          |             |                          |  |
|              | vv          | vt          | vv                    | vt          |                          |  |
| 1            | √           |             |                       |             |                          |  |
| 2            | √           | √           |                       |             |                          |  |
| 3            | √           | √           |                       |             |                          |  |
| 4            | √           | √           | √                     |             |                          |  |
| 5            | √           | √           | √                     | √           |                          |  |
| 6            | √           | √           | √                     | √           | √                        |  |
| 7            | √           | √           | √                     | √           | √                        |  |
| Cholec80     |             |             |                       |             |                          |  |
| Shot         | 1           | 4           |                       | 8           |                          |  |
| 1            | 35.66/28.06 | 38.88/31.92 |                       | 40.67/33.74 |                          |  |
| 2            | 37.60/30.30 | 43.56/36.02 |                       | 45.84/38.17 |                          |  |
| 3            | 41.88/36.16 | 43.86/37.87 |                       | 44.72/38.94 |                          |  |
| 4            | 45.84/36.97 | 47.68/38.58 |                       | 49.03/39.80 |                          |  |
| 5            | 48.91/38.32 | 47.85/39.91 |                       | 49.93/41.29 |                          |  |
| 6            | 48.15/39.25 | 49.44/40.21 |                       | 49.95/40.97 |                          |  |
| 7            | 48.35/39.36 | 49.31/40.48 |                       | 50.59/41.37 |                          |  |
| BernBypass70 |             |             |                       |             |                          |  |
| Shot         | 1           | 4           |                       | 8           |                          |  |
| 1            | 21.97/15.61 | 31.45/21.84 |                       | 34.09/23.42 |                          |  |
| 2            | 22.18/15.31 | 31.25/21.62 |                       | 33.87/23.18 |                          |  |
| 3            | 35.47/23.93 | 40.57/26.80 |                       | 41.92/27.53 |                          |  |
| 4            | 37.11/24.51 | 40.54/26.77 |                       | 41.51/27.43 |                          |  |
| 5            | 36.71/24.09 | 39.93/26.51 |                       | 40.49/27.12 |                          |  |
| 6            | 40.15/25.80 | 41.60/27.02 |                       | 42.24/27.71 |                          |  |
| 7            | 41.40/26.48 | 42.90/27.73 |                       | 43.18/28.21 |                          |  |

Note: Values are reported as Acc%/F1%. Acc, accuracy; F1, F1 score; SurgVLP, Surgical Vision–Language Pre-training; Atomic-Proto, atomic-level prototype; vv, video-video matching; vt, video-text matching. √ indicates that the module is used. Shot 0 indicates zero-shot transfer.

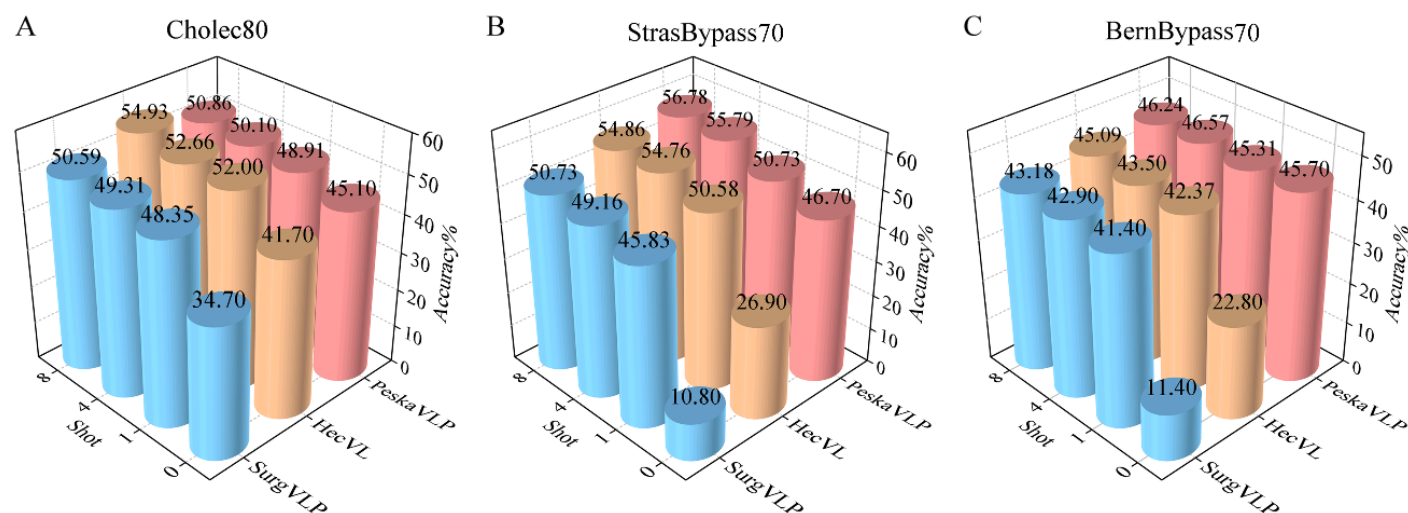
ously calculate similarities with both prototypes and obtain the final results through weighted fusion. Two-Stage Few-Shot Adaptation focuses on the performance degradation of VLMs when facing new categories and proposes a two-stage adaptation method [25]. The first stage learns a feature extractor that generalizes well to both base and new categories, while the second stage trains a linear classifier, achieving significant improvements over zero-shot transfer.

Across all datasets and shot settings, ALF-Surg consistently achieves the best performance, establishing a new SOTA for few-shot SPR. Our model leverages LLM-derived atomic-level procedural descriptions and constructs multi-level prototypes. The proposed multi-level multimodal matching enables the model to jointly utilize fine-grained atomic cues, global phase-level visual context, and structured video-text correspondence. Specifically, compared with the zero-shot transfer performance of the base model SurgVLP, ALF-Surg achieves significant improvements on all three datasets [11]. With only one shot, the

model achieves improvements of 13.65%/14.96% in accuracy and F1 score on Cholec80 (35.03%/29.68% on StrasBypass70 and 30.00%/19.28% on BernBypass70). Compared with other few-shot methods, ALF-Surg requires only a small number of labeled samples (1-shot), achieving a performance improvement of 12.44%/7.15% over Two-Stage Few-Shot Adaptation, the second-best method on Cholec80 [25]. On the more challenging BernBypass70 dataset, ALF-Surg improves over Proto-CLIP, the second-best method on this dataset, by 13.52%/1.55% [24].

### 3.2 Analysis of network components

To better understand the contribution of each module in ALF-Surg, we conduct comprehensive ablation studies on the Cholec80 and BernBypass70 datasets. We progressively disable or enable different components in prototype construction and multimodal matching, forming seven variant configurations, as shown in **Table 2**. These ablation experiments pri-



**Figure 5. Comparative experiments on the impact of different VLMs on three datasets.** Values are reported as accuracy (%). VLM, vision-language model; SurgVLP, Surgical Vision-Language Pre-training; HecVL, Hierarchical Video-Language Pretraining; PeskaVLP, Procedure-Aware Surgical Video-Language Pretraining with Hierarchical Knowledge Augmentation.

marily evaluate the role of atomic-level prototypes, the effectiveness of multi-level matching, the contribution of text alignment, and the importance of the mutual exclusion loss.

When class-level prototypes are constructed solely from visual blocks without any atomic semantic guidance (Settings 1 and 2), the model shows only modest improvements on Cholec80 but more noticeable gains on BernBypass70 compared with the zero-shot SurgVLP baseline. Nevertheless, the performance remains lower than that of settings incorporating atomic semantic guidance, indicating that holistic visual cues alone are insufficient for robust recognition across datasets. This indicates that, with a small number of samples, relying solely on holistic visual cues is insufficient to distinguish different surgical phases. Introducing atomic semantic information significantly improves model performance (Setting 3), validating the necessity of decomposing high-level surgical phases into LLM-derived atomic steps. Atomic-level representations explicitly encode “what to do”, “how to do it”, and “what tools to use”, providing fine-grained procedural cues and significantly enhancing prototype distinguishability. When combined with atomic-level matching (Settings 5 and 6), the robustness of recognition is further enhanced. The query video is matched not only with the overall representation but also with atomic operations, thereby enabling the differentiation of visually similar but functionally different phases. Adding video-text matching as an auxiliary alignment constraint generally improves or maintains competitive performance in most comparisons (Settings 1–2, 3–4, and 5–6), although the gains are not uniform across all datasets, shot settings, and metrics. Similarly, the transition from Setting 6 to Setting 7 generally supports the usefulness of the mutual exclusion loss, while some individual indicators show limited or no improvement. By encouraging different atomic concepts to attend to distinct temporal regions, the model avoids collapsing multiple atomic descriptions onto

the same frames. Nevertheless, its effectiveness may be limited in complex scenarios in which atomic actions heavily overlap or lack clear visual boundaries.

Overall, the ablation experiments demonstrate that atomic-level semantics are crucial for modeling fine-grained differences in surgical procedures. Multi-level multimodal matching leverages the complementary advantages of atomic-level evidence and global visual context, forming a cohesive system in conjunction with the mutual exclusion loss. This significantly improves model transferability to different institutions and surgical procedures using only a small number of labeled samples.

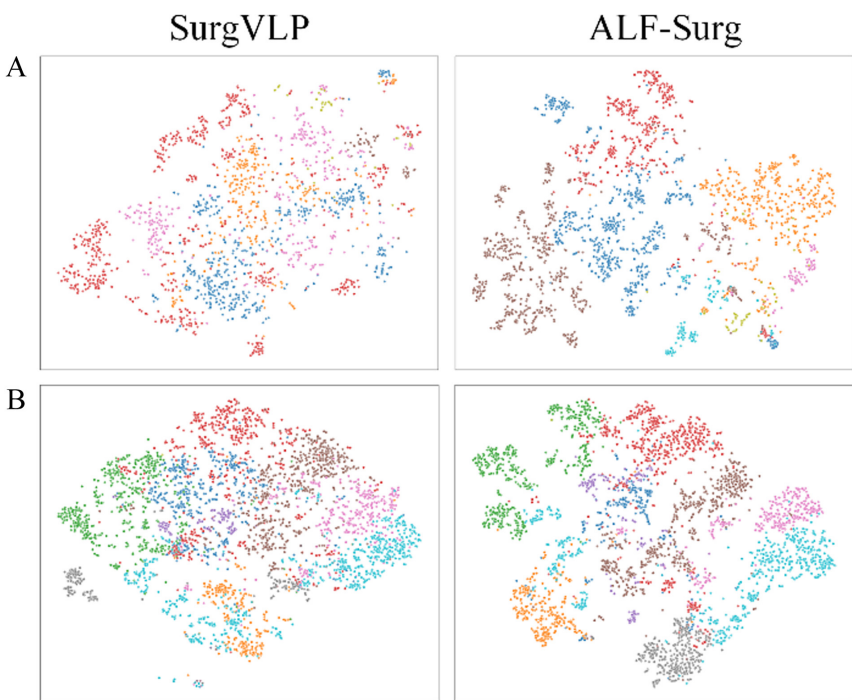
### 3.3 Analysis of pre-trained VLMs

To verify the effectiveness of ALF-Surg, we further investigate the performance of different pre-trained VLMs, including SurgVLP, Hierarchical Video-Language Pretraining (HecVL), and Procedure-Aware Surgical Video-Language Pretraining with Hierarchical Knowledge Augmentation (PeskaVLP), as frozen encoders on the few-shot SPR task [26, 27]. As shown in **Figure 5**, PeskaVLP achieves the strongest zero-shot performance, followed by HecVL and SurgVLP. By replacing the frozen encoder and retraining the model, ALF-Surg further improves its recognition accuracy on the three datasets. Replacing SurgVLP with HecVL yields substantial performance: 54.93% on Cholec80, 54.86% on StrasBypass70, and 45.09% on BernBypass70 in the 8-shot setting. Compared with HecVL zero-shot transfer, these results correspond to gains of +13.23% (Cholec80), +27.96% (StrasBypass70), and +22.29% (BernBypass70). Interestingly, the improvements obtained with PeskaVLP are more moderate than those obtained with HecVL. This is expected, as PeskaVLP models the temporal

**Table 3. Performance comparison using different numbers of input video frames**

| Frames | Shots | Cholec80 Acc%/F1% | StrasBypass70 Acc%/F1% | BernBypass70 Acc%/F1% |
|--------|-------|-------------------|------------------------|-----------------------|
| 4      | 1     | 45.31/38.25       | 47.89/39.01            | 41.30/27.46           |
|        | 4     | 48.35/41.30       | 50.38/40.55            | 44.56/29.35           |
|        | 8     | 49.57/41.93       | 51.53/41.41            | 45.48/30.14           |
| 8      | 1     | 48.35/39.36       | 45.83/36.58            | 41.40/26.48           |
|        | 4     | 49.31/40.48       | 49.16/39.40            | 42.90/27.73           |
|        | 8     | 50.59/41.37       | 50.73/40.43            | 43.18/28.21           |
| 16     | 1     | 46.87/37.87       | 41.27/34.47            | 39.24/24.85           |
|        | 4     | 49.10/39.72       | 44.12/35.76            | 39.86/26.38           |
|        | 8     | 50.22/40.41       | 45.27/36.47            | 39.68/26.44           |

Note: Values are reported as Acc%/F1%. Acc, accuracy; F1, F1 score. Frame indicates the number of consecutive input frames per sample. Shot indicates the number of support samples per class.



**Figure 6. Visualization of the t-SNE distributions of SurgVLP and ALF-Surg on the test sets of the Cholec80 and StrasBypass70 datasets. (A) Cholec80; (B) StrasBypass70. t-SNE, t-distributed stochastic neighbor embedding; SurgVLP, Surgical Vision–Language Pre-training; ALF-Surg, Atomic Language-Guided Few-Shot Surgical Phase Recognition.**

hierarchy during pre-training and is trained using a large number of multi-scale video-text pairs, similar to ALF-Surg.

3.4 Analysis of the number of input video frames

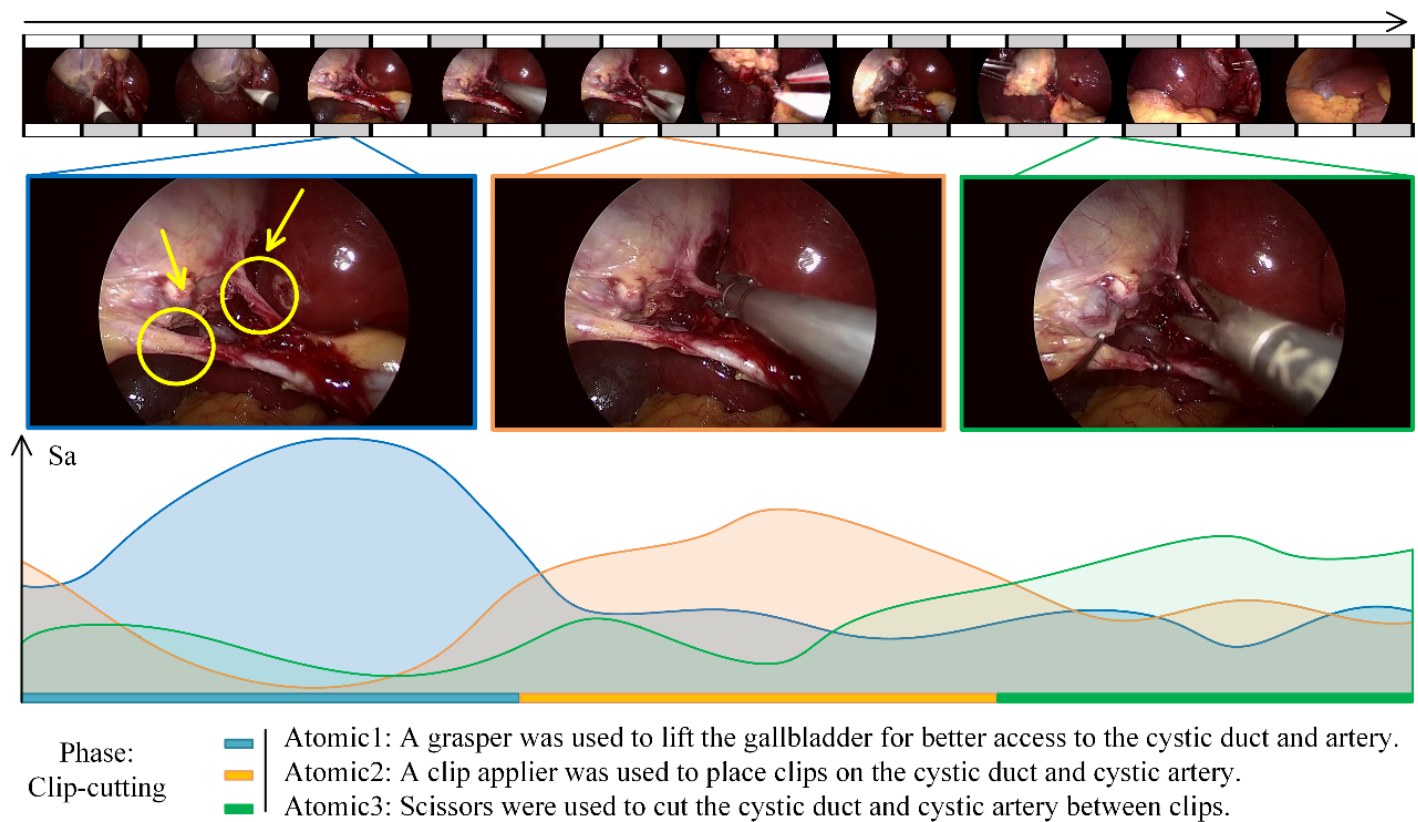
ALF-Surg treats eight consecutive frames as a single sample by default. To further understand the impact of temporal granularity, we conduct experiments using video clips of four and 16 frames. As shown in **Table 3**, the model still maintains good performance when the video clip length is reduced to four

frames. Conversely, increasing it to 16 frames leads to a performance drop, especially on the more challenging StrasBypass70 and BernBypass70 datasets. ALF-Surg uses atomic-level text descriptions, and these operations are typically completed within a very short time window. Therefore, even short clips contain enough information to enable reliable atomic-level matching, prototype alignment, and other processes. Extending the length of video clips to 16 frames or longer can result in excessive temporal coverage. In this case, unique atomic cues are diluted by irrelevant or mixed contextual information, reducing the focus on frames that truly express atomic semantics and thus degrading model performance.

4 DISCUSSION

4.1 Visualization of prototype similarity

To further analyze our model, we visualize the embedding spaces learned by the baseline model (SurgVLP) and our method on the Cholec80 and StrasBypass70 datasets. As illustrated in **Figure 6**, ALF-Surg significantly enhances feature discrimination across surgical phases. First, it should be noted that, due to differences in surgical complexity and anatomical exposure, the duration of each phase varies greatly, resulting in significant inter-class imbalance in the data distribution. Benefiting from atomic-level language guidance and multi-level prototype construction, ALF-Surg effectively encodes distinct functional semantics embedded in each phase and pulls query representations closer to their corresponding prototypes. Consequently, the learned feature manifolds demonstrate more compact intra-class clustering and clearer inter-class margins, confirming enhanced



**Figure 7. Visualization of atomic-level matching.** The yellow arrows and circles indicate the exposed cystic duct and cystic artery.  $S_a$  denotes the atomic-level video-video matching score.

discriminative ability and knowledge adaptation to the downstream task.

#### 4.2 Visualization of atomic-level matching

To further demonstrate ALF-Surg's ability to perceive atomic-level operations in different phases, we visualize the atomic-level video-video matching results. In **Figure 7**, we present an example from the clip-cutting phase of laparoscopic cholecystectomy. Clinically, this stage requires first using clips to close the exposed cystic duct and cystic artery to prevent bile and other fluids from leaking out, followed by cutting to facilitate subsequent gallbladder removal. The visualization demonstrates that ALF-Surg progressively retrieves these atomic operations in temporal order through explicit alignment between atomic-level prototypes and frame-level features. This indicates that the model successfully captures fine-grained interaction semantics rather than relying solely on global context.

#### 4.3 Prompt design

To fully leverage the rich prior knowledge embedded in LLMs, such as GPT-4o, we convert high-level surgical phase names into fine-grained, temporally ordered atomic action descriptions. Specifically, to mitigate potential ambiguity and

temporal overlap, we enforce a strict “temporal order” rule and limit the output to “visual-only actions”. This guides the LLM to produce distinct, non-overlapping steps that strictly follow the surgical workflow, thereby reducing semantic confusion during the alignment process. The complete prompt used in our work is provided below:

**System Role:** You are a medical domain expert specialized in laparoscopic cholecystectomy. You will generate detailed, visual, and temporally ordered atomic action descriptions for each surgical phase in the Cholec80 dataset.

**Task:** Given a phase label and its corresponding frequently used instruments and anatomical targets, decompose the phase into an ordered list of 2–4 atomic action descriptions that reflect realistic laparoscopic surgical actions.

**Rules:** Temporal order: Write atomic actions from the beginning to the end of the phase. Visual-only actions: Describe only what can be directly observed in the laparoscopic video feed. Include instruments and targets: Every atomic action must mention the relevant instruments and targets. No unnecessary details: Omit texture, color, or emotional descriptions unless visually critical. Concise and precise: Each atomic action should contain fewer than 20 words. Output Format: “Phase Label”: “<phase name>”, “Atomic Actions”: [“<first

atomic action description>”, “<second atomic action description>”, “<third atomic action description>”].

## 5 CONCLUSION

In summary, we present ALF-Surg, a lightweight framework designed to enhance the few-shot transferability of vision–language models for SPR. By harnessing the medical knowledge within LLMs to generate atomic-level action descriptions and by employing a multi-level matching strategy, our method robustly captures complex spatiotemporal dynamics and instrument-tissue interactions. This design effectively bridges the semantic gap present in conventional supervision, enabling the model to generalize well even with minimal annotated data. Extensive evaluations on three multi-center datasets (Cholec80, StrasBypass70, and BernBypass70) confirm that ALF-Surg significantly surpasses zero-shot baselines and establishes new SOTA performance in few-shot settings. By enabling rapid model adaptation with low annotation costs, our framework provides a scalable path toward real-world deployment and holds great promise for the future of computer-assisted intervention.

## DECLARATIONS

### Author contributions

Houlong He, Chengli Song, and Lin Mao conceived this project. Houlong He performed the experiments. Houlong He, Chengli Song, and Lin Mao helped perform the analysis with constructive discussions. All authors contributed to the writing of the manuscript.

### Funding

The authors confirm that there is no funding or financial support to report for this article.

### Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

All authors acknowledge and agree to submit this article to the journal Progress in Medical Devices.

### Competing interests

The authors declare that they have no competing interests.

### Acknowledgements

Not applicable.

## REFERENCES

- [1] Lin C, Zhu Z, Zhao Y, Zhang Y, He K, Zhao Y. SGT++: Improved scene graph-guided transformer for surgical report generation. *IEEE Trans Med Imaging*. 2024 Apr;43(4):1337-1346. <https://doi.org/10.1109/TMI.2023.3335909>
- [2] Khanna A, Wolf T, Frank I, Krueger A, Shah P, Sharma V, et al. Enhancing accuracy of operative reports with automated artificial intelligence analysis of surgical video. *J Am Coll Surg*. 2025 May 1;240(5):739-746. <https://doi.org/10.1097/XCS.0000000000001352>
- [3] Wang Z, Liu C, Zhang S, Dou Q. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: *Proc Int Conf Med Image Comput Comput-Assist Intervent*. Springer; 2023. p. 101-111. [https://doi.org/10.1007/978-3-031-43996-4\\_10](https://doi.org/10.1007/978-3-031-43996-4_10)
- [4] Chen T, Yuan K, Srivastav V, Navab N, Padoy N. Text-driven adaptation of foundation models for few-shot surgical workflow analysis. *Int J Comput Assist Radiol Surg*. 2025 Jun;20(6):1175-1183. <https://doi.org/10.1007/s11548-025-03341-0>
- [5] Wu Y, Lu Y, Zhou Y, Ding Y, Liu J, Ruan T. MKGF: A multi-modal knowledge graph based RAG framework to enhance LVLMS for medical visual question answering. *Neurocomputing*. 2025 Jun 28;635:129999. <https://doi.org/10.1016/j.neucom.2025.129999>
- [6] Ramesh S, Dall’Alba D, Gonzalez C, Yu T, Mascagni P, Mutter D, et al. Weakly supervised temporal convolutional networks for fine-grained surgical activity recognition. *IEEE Trans Med Imaging*. 2023 Sep;42(9):2592-2602. <https://doi.org/10.1109/TMI.2023.3262847>
- [7] Li Y, Zhao G, Li C, Shi W, Jiang Z, Zhang Z, et al. STSANet: Spatial temporal-self-aggregation network for surgical phase recognition. *Inf Fusion*. 2026 Feb;126 Part B:103646. <https://doi.org/10.1016/j.inffus.2025.103646>
- [8] Qin Z, Yi H, Lao Q, Li K. Medical image understanding with pretrained vision language models: A comprehensive study. In: *The Eleventh International Conference on Learning Representations*; 2023 May 1-5; Kigali, Rwanda. ICLR; 2023.
- [9] Dissanayake T, George Y, Mahapatra D, Sridharan S, Fookes C, Ge Z. Few-shot learning for medical image segmentation: A review and comparative study. *ACM Comput Surv*. 2025;58(1):1-36. <https://doi.org/10.1145/3746224>
- [10] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning*; 2021 Jul 18-24; Virtual. PMLR; 2021. p. 8748-8763.
- [11] Yuan K, Srivastav V, Yu T, Lavanchy JL, Marescaux J, Mascagni P, et al. Learning multi-modal representations by watching hundreds of surgical video lectures. *Med Image Anal*. 2025 Oct;105:103644. <https://doi.org/10.1016/j.media.2025.103644>
- [12] Hu M, Yuan K, Shen Y, Tang F, Xu X, Zhou L, et al. OphCLIP: Hierarchical retrieval-augmented learning for ophthalmic surgical video-language pretraining. In: *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2025 Oct 19-25;

- Honolulu, HI, USA. IEEE; 2026. p. 1-12. <https://doi.org/10.1109/ICCV51701.2025.01845>
- [13] Quan H, Li X, Hu D, Nan T, Cui X. Dual-channel prototype network for few-shot pathology image classification. *IEEE J Biomed Health Inform.* 2024 Jul;28(7):4132-4144. <https://doi.org/10.1109/JBHI.2024.3386197>
- [14] Wang C, Liu F, Chen Y, Frazer H, Carneiro G. Cross- and intra-image prototypical learning for multi-label disease diagnosis and interpretation. *IEEE Trans Med Imaging.* 2025 Jun; 44(6):2568-2580. <https://doi.org/10.1109/TMI.2025.3541830>
- [15] Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, et al. Clip-adapter: Better vision-language models with feature adapters. *Int J Comput Vis.* 2024;132(2):581-595. <https://doi.org/10.1007/s11263-023-01891-x>
- [16] Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. *Int J Comput Vis.* 2022;130(9):2337-2348. <https://doi.org/10.1007/s11263-022-01653-1>
- [17] Shu M, Nie W, Huang D, Yu Z, Goldstein T, Anandkumar A, et al. Test-time prompt tuning for zero-shot generalization in vision-language models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. *Advances in Neural Information Processing Systems 35. 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*; 2022 Nov 28–Dec 9; New Orleans, LA. Red Hook (NY): Curran Associates, Inc.; 2022. p. 14274-14289. <https://doi.org/10.52202/068431-1038>
- [18] Qian Z, Yao X, Huang Y, Zhang C, Ying J, Sun H. Beyond label semantics: Language-guided action anatomy for few-shot action recognition. In: *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2025 October 19-25; Honolulu, HI, USA. IEEE; 2026. p. 10421-10431. <https://doi.org/10.1109/ICCV51701.2025.00970>
- [19] Shi Y, Wu X, Lin H, Luo J. Commonsense knowledge prompting for few-shot action recognition in videos. *IEEE Trans Multimedia.* 2024;26:8395-8405. <https://doi.org/10.1109/TMM.2024.3361157>
- [20] Zhang R, Zhang W, Fang R, Gao P, Li K, Dai J, et al. Tip-adapter: Training-free adaption of clip for few-shot classification. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. *Computer Vision – ECCV 2022: 17th European Conference on Computer Vision*; 2022 Oct 23-27; Tel Aviv, Israel. Springer; 2022. p. 493-510. [https://doi.org/10.1007/978-3-031-19833-5\\_29](https://doi.org/10.1007/978-3-031-19833-5_29)
- [21] Shakeri F, Huang Y, Silva-Rodríguez J, Bahig H, Tang A, Dolz J, et al. Few-shot adaptation of medical vision-language models. In: *Proc Int Conf Med Image Comput Comput-Assist Interv.* Springer; 2024. p. 553-563. [https://doi.org/10.1007/978-3-031-72390-2\\_52](https://doi.org/10.1007/978-3-031-72390-2_52)
- [22] Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging.* 2017;36(1):86-97. <https://doi.org/10.1109/TMI.2016.2593957>
- [23] Lavanchy JL, Ramesh S, Dall’Alba D, Gonzalez C, Fiorini P, Müller-Stich BP, et al. Challenges in multi-centric generalization: phase and step recognition in Roux-en-Y gastric bypass surgery. *Int J Comput Assist Radiol Surg.* 2024 Nov;19(11):2249-2257. <https://doi.org/10.1007/s11548-024-03166-3>
- [24] Palanisamy K, Chao Y, Du X, Xiang Y. Proto-CLIP: Vision-language prototypical network for few-shot learning. In: *Proc IEEE/RSJ Int Conf Intell Robots Syst.* IEEE; 2024. p. 2594-2601. <https://doi.org/10.1109/IROS58592.2024.10801660>
- [25] Farina M, Mancini M, Iacca G, Ricci E. Rethinking few-shot adaptation of vision-language models in two stages. In: *Proc IEEE Conf Comput Vis Pattern Recog.* 2025. p. 29989-29998. <https://doi.org/10.1109/CVPR52734.2025.02791>
- [26] Yuan K, Srivastav V, Navab N, Padoy N. HecVL: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: *Proc Int Conf Med Image Comput Comput-Assist Interv.* Springer; 2024. p. 306-316. [https://doi.org/10.1007/978-3-031-72089-5\\_29](https://doi.org/10.1007/978-3-031-72089-5_29)
- [27] Yuan K, Srivastav V, Navab N, Padoy N. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In: *Advances in Neural Information Processing Systems 37*; 2024 Dec 10-15; Vancouver, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2024. p. 122952-122983. <https://doi.org/10.52202/079017-3907>