Universiteit Leiden
The Netherlands

Data Science and Artificial Intelligence

Design and Evaluation of an Active Vision System for Surgical Instrument Detection and State Classification

Adrien Joon-Ha Im

Supervisors:
Dr. Daan Pelt & Dr. Jiayang Shi

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

16/06/2026

Abstract

Recognizing handheld surgical instruments reliably is a challenging computer vision problem, particularly in open surgery where instruments are visually similar and conditions vary. This thesis presents the design and evaluation of an active multi-view vision system for surgical instrument recognition in a controlled setting, addressing two tasks: category recognition across five instrument classes (clamps, needle holders, scalpels, shears, and tweezers) and binary hinge-state classification (open or closed) for hinged instruments. A dedicated dataset of 2,520 images was collected under controlled variation in viewpoint, background, and lighting. Trained vision-only models achieved strong performance: the best object detector (YOLO26n) reached 98.2% mAP@50:95 and the best hinge-state classifier (EfficientNet-B0) reached 98.1% accuracy. Zero-shot vision-language models were evaluated as a comparative baseline but remained substantially weaker, reaching only 72.2% category accuracy and 62.4% hinge-state accuracy, with considerably higher inference latency. A key finding is that the two recognition tasks rely on different visual cues and benefit from different viewpoints: close-up views were most informative for category recognition, where fine-grained local geometry such as jaw shape and tip structure is discriminative, while top-down views were most effective for hinge-state classification, where the global separation between instrument branches determines the correct label. This asymmetry motivates adaptive viewpoint selection. An active acquisition pipeline was therefore evaluated, requesting additional views only when prediction confidence was insufficient. The strongest active configuration improved end-to-end accuracy on hinged instrument scenes from 55.0% using only the initial top-down view to 73.3%, exceeding the corresponding full three-view fusion result of 70.0%, while requiring on average only 2.50 views per scene. These results demonstrate that active multi-view acquisition with trained vision-only models supports accurate and efficient surgical instrument recognition under controlled conditions, and that viewpoint selection should be treated as task-specific rather than task-agnostic.

Contents

1 Introduction

Modern medicine increasingly relies on artificial intelligence (AI) and computer vision (CV), particularly as automated and computer-assisted procedures become more common in the operating room. In surgery, computer vision can be used to analyze visual information from the surgical scene and support a better understanding of what is happening during an operation. This can include the recognition of anatomical structures, surgical phases, and surgical instruments. By extracting clinically relevant information from surgical video, computer vision systems can contribute to making surgical procedures more measurable, more efficient, and ultimately safer [MAS+22].

One important aspect of surgical scene understanding is instrument detection, which consists of localizing surgical instruments with bounding boxes or segmentation masks and classifying them. Reliable information about the instrument in a surgical scene supports workflow analysis, such as identifying which part of the surgery is being performed, and quantifying how much time is spent on specific surgical tasks. This information can in turn support clinical applications such as skill assessment, the development of supporting tools that give intraoperative guidance and feedback to surgeons, or more automation in the operating room with robotics [FHKS22, MAS+22, XLC+25].

Despite recent advances in surgical computer vision, reliable real-time recognition of surgical instruments in open surgery remains a challenging computer vision problem [MAS+22, FHKS22]. Many existing computer vision systems have been developed for minimally invasive surgery, where cameras and specialized instruments are inserted into the body via a small opening and provide a relatively controlled view of the surgical scene [RRP24]. In contrast, open surgery takes place in a less controlled visual environment, where images involve greater variation in viewpoint, illumination, background, and hand-instrument interactions. In addition, occlusions caused by the surgeon's hands further increase the difficulty of visual recognition [SHK+21, XLC+25]. Handheld general surgical instruments that are used in open surgery also pose a fine-grained recognition challenge as many of them are made of similar metallic materials and often share a broadly similar overall shape, despite serving different clinical functions. As a result, distinguishing between instrument categories may depend on subtle visual cues, particularly for visually similar tools such as clamps and needle holders [LCHW21].

In addition to recognizing instrument category, some applications may also require information about the functional state of the instrument. For hinged instruments such as clamps or surgical scissors, this includes determining whether the instrument is open or closed. Such state information may be relevant in the understanding of how an instrument is being used during a procedure. State recognition can be more difficult than object recognition alone, as different object categories share the same state while the same object may appear in multiple configurations [GPAP21]. As a result, a surgical vision system must not only distinguish between different tool categories, but also detect subtle geometric differences such as hinge angle or jaw separation, introducing an additional layer of complexity to the recognition task.

This thesis presents the design and evaluation of an active vision system for surgical instrument recognition. The study focuses on five categories of handheld general surgical instruments: clamps, needle holders, scalpels, shears, and tweezers [LS22]. Two related computer vision tasks are addressed. The first task is instrument category recognition, formulated as an object detection problem in

which the system must localize the instrument in the image and assign it to the correct category. The second task is hinge-state classification for hinged instruments, formulated as a binary classification problem in which the system predicts whether the detected instrument is open or closed.

The core experiments in this thesis are conducted using a dedicated dataset collected for this study under a closed-world assumption, in which the instrument categories and hinge-state labels are predefined. In order to evaluate recognition performance under variation, the dataset includes multiple backgrounds, lighting conditions, and predefined viewpoints. Furthermore, a limited generalization experiment is performed using previously unseen instrument instances belonging to the same predefined categories, allowing the comparison between performance on known objects and robustness to new instances within the same class.

Beyond recognition performance alone, this thesis investigates how viewpoint selection influences instrument recognition performance. Prior work has shown that multi-view imaging can improve robustness by reducing the impact of occlusion and loss of visibility from a single camera perspective [BGPL22]. To build on this, the thesis examines whether an adaptive viewpoint selection strategy can improve recognition performance while reducing the number of acquired views. To address these questions, multiple vision-only models based on convolutional neural networks (CNNs) are evaluated for both instrument category recognition and hinge-state classification. In addition, several vision-language models (VLMs) are assessed in a zero-shot setting to compare specialized vision-only models with more general multimodal approaches that combine both visual and textual capabilities [ZAW+23].

Research Questions The main research question addressed in this thesis is the following:

> **Main RQ** How can an active multi-view imaging system be designed to reliably recognize the category and hinge state of handheld surgical instruments under variation in viewpoint, lighting, and background?

To answer this question, the thesis investigates which viewpoints provide the most discriminative information for instrument category recognition and hinge-state classification, how accurately vision-only models can perform these tasks on the collected dataset, whether adaptive multi-view acquisition improves recognition performance compared with single-view acquisition, and how vision-language models compare with specialized vision-only models on the same tasks. These sub-questions are formalized as:

RQ1 Which viewpoints and visual regions provide the most discriminative information for surgical instrument category recognition and hinge-state classification?

RQ2 How accurately can the proposed vision-only system recognize instrument category and hinge state across variations in viewpoint, lighting, and background?

RQ3 To what extent does an adaptive multi-view acquisition strategy improve recognition performance compared with single-view recognition and fusion of all available views?

RQ4 How do zero-shot vision-language models compare with specialized vision-only models for instrument category recognition and hinge-state classification?

Thesis Overview Section 2 presents the background and related work relevant to this thesis. It introduces the handheld surgical instruments and functional state labels considered in this study, and discusses the main computer vision concepts used in the thesis, including vision-only models, vision-language models, viewpoint selection, and multi-view perception. Section 3 describes the methodology, including dataset collection, annotation, experimental design, model training, and pipeline development. Section 4 presents the experimental results, and Section 5 interprets and discusses these findings. Finally, Section 6 concludes the thesis and outlines directions for future research. This bachelor thesis was supervised by Dr. Daan Pelt and Dr. Jiayang Shi at the Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University.

2 Background and Related Work

This chapter introduces the clinical and technical background relevant to this thesis. It first defines surgical instrument categories and state labels considered in this study. It then outlines modern deep learning techniques for visual recognition, before reviewing prior work on surgical instrument recognition using computer vision, especially in relation to open surgery settings, instrument state recognition, multi-view perception, and vision-language approaches. Finally, the chapter identifies the research gap addressed by this thesis.

2.1 Surgical Instruments

A precise definition of the instrument categories is necessary because surgical instruments can be described at different levels of granularity. In clinical practice, instruments are often differentiated by exact subtype, geometry, and use. In this thesis, however, the recognition task is formulated at a broader level of categorization than at the level of exact product names. This choice was made because the primary focus of this thesis is on the image acquisition strategy for vision systems, rather than on a performance analysis of computer vision models on fine-grained object categories.

The study focuses on five categories of common handheld surgical instruments: scalpels, tweezers, shears, clamps, and needle holders, following the general categorization described by Liehn Schlautmann [LS22]. These categories were selected as they represent instruments that are widely used across procedures and medical specialties. The chosen instruments also span both visually similar and visually distinct categories. For some instruments, the category labels used in this thesis are simplified. For example, the Adson tissue forceps included in the dataset are assigned to the broader category “tweezers”.

(a) Scalpel (b) Tweezers (c) Shears (d) Clamp (e) Needle holder

Figure 1. Five surgical instrument categories used in this study.

Scalpel Scalpels are cutting instruments used primarily for making incisions. They typically consist of a handle and a blade, although both handle design and blade shape may vary across instruments and manufacturers [LS22]. Within this study, only a scalpel handle is used and is treated as a non-hinged category.

Tweezers Tweezers, also referred to clinically as forceps, are non-hinged instruments used to hold or manipulate tissue. They consist of two spring-loaded halves that close when pressure is applied and reopen when released. Unlike clamps or needle holders, they do not include finger rings, a hinge joint, or a ratchet locking mechanism, which makes them visually distinct from the hinged instrument categories considered in this study [LS22].

Shears Shears, or surgical scissors, are hinged cutting instruments used for cutting and dissection. Their blades can vary in length, curvature, and profile depending on the required depth and anatomy. Shears consist of two articulated halves connected to a joint, with a handle section and a cutting section [LS22].

Clamp Clamps are hinged grasping instruments that are used to grip tissue or material with both jaws. They exist in a variety of shapes and profiles, including straight, curved, and bayonet-shaped variants. Most clamps include ring handles and a ratchet locking mechanism that allows the jaws to remain closed until released by the surgeon [LS22].

Needle holder Needle holders visually resemble clamps in overall appearance and construction, sharing ring handles and a hinge. However, needle holders are specifically designed to exclusively hold and guide a surgical needle. The coarseness of the jaws is adjusted based on the size of the needle, and each needle holder has a corresponding needle size [LS22].

Among the categories selected for this study, the clamp and needle holder form the most visually similar pair. Both have ring handles, are hinged, include a locking mechanism, and have narrow tips. Their distinction therefore depends less on global shape and more on localized geometric features. This makes them a particularly interesting test case for computer vision, as successful recognition may require attention to fine-grained visual elements rather than the overall aspect of the instrument.

2.1.1 State Definition

In addition to instrument categories, this study also considers two functional states of the hinged instruments: *open* and *closed*. This applies to only hinged instruments: needle holders, clamps, and shears. For non-hinged instruments, the dataset label `NA` is used. For clamps and needle holders, the instrument is considered closed when the jaws are closed, and the ratchet mechanism is in the locked position. For shears, the instrument is considered closed when the two handles are in contact. Any visible separation between the two articulated parts is labeled as open. This definition allows the state recognition task to be applied consistently across the hinged instrument categories.

2.2 Deep Learning for Visual Recognition

Modern computer vision systems are commonly based on deep learning. Deep learning has become a central technique in computer vision, particularly for recognition tasks such as image classification, object detection, and segmentation [Sze22].

This thesis considers two main types of vision models: vision-only models and vision-language models (VLMs). Vision-only models take an image as input and output visual predictions such as class labels or bounding boxes. These models learn visual features from annotated images, and usually are trained or fine-tuned for specific tasks. Convolutional neural networks (CNNs), for example, process images using convolutional filters that are applied over an image. This allows the neural network to detect local visual patterns while preserving spatial structure [Sze22].

For image classification, the model assigns a single class label to an input image. Architectures such as ResNet and EfficientNet are commonly used for this type of task [HZRS16, TL19]. ResNet introduced residual connections, which help train deeper networks by allowing information to pass more directly through the model [HZRS16]. EfficientNet uses a compound scaling strategy to balance network depth, width, and input resolution, making it efficient for classification tasks [TL19].

Object detection is an extension of image classification, as it requires the model to both classify and localize objects in an image. Instead of producing a single image label, an object detector predicts bounding boxes, class labels, and confidence scores for detected objects [Sze22]. YOLO, short for "You Only Look Once", is a single-stage object detector that predicts object locations and class probabilities in one forward pass through the network [RDGF16]. This makes YOLO suitable for applications where inference speed is important, such as interactive or active vision systems.

A common strategy in modern deep learning is transfer learning [Sze22]. Instead of training a model entirely from scratch, models are initialized using weights that are pretrained on large public research datasets such as ImageNet for image classification or COCO (Common Objects in Context) for object detection [DDS+09, LMB+14]. These pretrained models have learned general visual features such as edges, textures, and shapes. During fine-tuning, the model parameters are then adapted to the target dataset. This is especially useful for computer vision in specialized domains, such as surgical images, where the datasets are often much smaller than general-purpose image datasets [Sze22].

The second type of model considered in this thesis is the vision-language model (VLM). Unlike vision-only models, VLMs are trained to connect visual information with natural language. Modern VLMs typically combine a visual encoder with a language model so that image features can be interpreted together with a text prompt [LWD+25, LYP+24]. This enables them to perform tasks such as image understanding, visual question answering, and zero-shot classification [RKH+21].

This distinction between vision-only models and VLMs is important for this thesis because they operate with different inputs and outputs. A vision-only detector such as YOLO receives an image as input and is optimized to output predefined class labels, bounding boxes, and confidence scores. In contrast, a VLM receives both an image and a text prompt, and outputs a natural language response. In this thesis, the vision-only models are trained or fine-tuned for the specific surgical instrument recognition tasks, whereas the VLMs are evaluated in a zero-shot setting using text prompts.

2.3 Computer Vision for Instrument Detection

Deep learning has been widely applied to surgical instrument detection, but most prior work has focused on minimally invasive surgery (MIS), cataract surgery, and other relatively constrained settings rather than on handheld instruments in open surgery [FHKS22]. In these environments, the camera viewpoint is often fixed, and the visual scene is more standardized. Open surgery, in contrast, presents a less constrained imaging setting, with greater variation in viewpoint, illumination, background, and hand-instrument interactions. Instruments may also be partially occluded by the surgeon's hands, which makes reliable recognition more difficult than in many MIS settings [FHKS22, SHK+21].

Within the literature on open surgery, there are two directions that can be distinguished. One focus of research is on egocentric imaging, or intraoperative video, where tools must be localized in the surgeon's field of view [SHK+21]. A second direction focuses on the recognition of tools on an instrument stand, or outside the body [KHE+25]. This second approach is more comparable to the present study since this thesis focuses on recognition of handheld instruments under controlled variation, rather than on full surgical-scene understanding.

Earlier approaches to surgical tool recognition included sensor-based systems and more classical image-based techniques, but recent work is dominated by deep learning methods. In particular, neural networks and modern object detectors have become the main approach in this task [FHKS22]. Fujii et al. evaluated several detection architectures for open surgery videos, including Faster R-CNN and RetinaNet with different backbones, and showed that surgical tool detection in open surgery is feasible while emphasizing the challenges associated with limited datasets and complex visual conditions [FHKS22]. More recent work has also shown strong performance with YOLO-based approaches [XLC+25]. In controlled settings, Lehr et al. also demonstrated that CNN-based recognition of surgical instruments can achieve high accuracy, although performance is obtained under more restricted conditions [LKB+23].

Accurate instrument detection is not only important because it identifies and localizes the tool, but also because it supports downstream tasks such as tracking, pose estimation, workflow analysis, remaining time prediction, and surgical skill assessment [FHKS22]. At the same time, the literature indicates that open-surgery detection remains challenging because large annotated datasets are relatively limited and because handheld instruments are often reflective, visually similar, and partially occluded [FHKS22, ZW17].

2.4 State Recognition in Computer Vision

State recognition refers to predicting the functional or physical configuration of an object rather than only identifying its category. In the case of surgical instruments, this is especially relevant for hinged tools such as clamps, needle holders, and shears, which can appear in open or closed configurations. Compared with category recognition, this task is more fine-grained because the object category may remain the same while only subtle geometric differences such as the hinge angle, or jaw separation determine the correct state label [GPAP21].

From a broader computer vision perspective, object state recognition is closely related to standard

object detection and classification, but poses a distinct challenge. Gouidis et al. distinguish object-state detection from standard object detection and suggest that state recognition is generally more difficult, partly because different object categories can share the same state, while the same object category can appear in multiple different states [GPAP21]. As a result, state recognition depends less on general object appearance, and more on localized visual cues.

State information is also closely connected to action and manipulation understanding. In many vision tasks, the state of an object is not only important as a static label, but also because it reflects how an object is being used or how it changes over time. Alayrac et al. [ASLLJ17] studied object states together with manipulation actions and state transitions, showing that recognizing object changes over time is an important aspect of understanding interactions in images and videos.

This is particularly relevant in surgical vision because the configuration of an instrument can describe its functional use rather than only its category. For example, a hinged instrument that is open or closed may indicate different stages of grasping, cutting, or manipulation. At the same time, recognizing such states can be visually challenging, because the distinction may depend on small geometric cues such as jaw separation, hinge angle, or handle position.

2.5 Viewpoint Selection and Multi-View Perception

Viewpoint selection is an important factor in visual recognition because different camera angles do not provide the same amount of visual information. For surgical instruments, this matters especially in open-surgery settings, where recognition can depend on local cues such as jaw shape, hinge angle, or tip geometry. A single viewpoint may hide these regions. This is particularly relevant when distinguishing between visually similar handheld instruments as the most informative parts are often small and can be partially occluded or poorly visible from a specific given angle [SHK+21, FHKS22].

Multi-view perception addresses this limitation by combining complementary information from several viewpoints. Basiev et al. showed this in an open-surgery setting with a multi camera system for tool classification. They argue that multi-camera setups can prevent failures caused by occlusion and loss of visibility from a single perspective. Their setup combines a top-view camera with a close-up view, giving a more complete picture of the surgical scene compared to a single camera setup [BGPL22].

Beyond surgical vision, recent work on multi-view understanding also supports the idea that additional views should not simply be added but selected. Hou et al. propose selecting the next most informative view for recognition in order to reduce computational cost and maintain strong performance. In other words, only a subset of the available views are necessary to achieve strong multi-view performance, rather than processing all views indiscriminately [HGZ24].

Taken together, these findings suggest that multi-view perception in computer vision is valuable not only because it adds redundancy but because different viewpoints reveal different discriminative features. For surgical instrument recognition, this motivates evaluating which views are the most informative for category recognition and for hinge-state classification. This study therefore examines this multi-view aspect of surgical instrument recognition in more detail.

2.6 Vision-Language Approaches

Vision-Language Models, or VLMs, are multimodal AI models that learn relationships between visual data and natural language, which allows them to process images together with text, and perform tasks such as answering questions about the image, captioning, and image classification [RKH+21, LWD+25]. One important strength of VLMs is that language can be used to describe the task, which makes it usable in settings where annotated data is missing, allowing them to be used in zero-shot setups [RKH+21].

In surgical settings, the use of VLM is relatively recent. Zhou et al. [ZAW+23], for example, showed that vision-language models can be used when prompting in surgical instrument segmentation tasks. At the same time, recent surveys show that there are important challenges that remain in specialized domains such as medicine. These challenges include domain shift, and difficulty when distinguishing fine-grained detail [LWD+25, SFB+26]. Such limitations are relevant in this thesis, as there are classification tasks between visually similar instruments, and also an additional task that involves distinguishing between subtle hinge-state differences.

For this reason, VLMs are included in this thesis as a comparative baseline against specialized vision-only models. Unlike the vision-only models, which are trained or fine-tuned directly for instrument detection and hinge-state classification, the VLMs are evaluated in a zero-shot setting using text prompts. This makes it possible to assess whether general-purpose multimodal VLMs can perform fine-grained surgical instrument recognition, and to what extent they can compete with specialized vision-only models in the context of fine-grained surgical instrument classification tasks.

2.7 Research Gap

Existing literature shows that surgical instrument recognition is feasible, but also highlights several limitations that are relevant for open surgery settings. First, most prior work has focused on MIS or other constrained environments rather than on handheld instruments in open surgery [FHKS22]. Second, prior studies often study instrument category recognition, whereas state recognition remains comparatively less explored [GPAP21]. Third, although multi-view and active perception setups have shown potential to improve robustness, there is still limited work that systematically examines how viewpoint selection and multi-view acquisition affect category recognition and hinge-state recognition for handheld instruments. In addition, while VLMs have recently been applied to surgical vision tasks [ZAW+23], there is limited literature on how the most recent models compare to vision-only models.

This thesis addresses these gaps by investigating recognition performance by viewpoint, comparing single-view and multi-view acquisition, and evaluating whether vision-language models offer a viable alternative to vision-only models.

3 Methodology

This chapter describes the methodology used to design and evaluate the proposed active vision system. It first explains the design, acquisition, annotation, and validation of the surgical instrument datasets. It then presents the vision-only models for instrument detection and hinge-state classification, including their training setup. Next, the unified pipeline, fusion method for multiple viewpoints, and threshold-based active view acquisition strategy are described. The chapter also introduces the zero-shot VLM setup, including the selected models, prompts, and output normalization. Finally, the evaluation protocol is defined for object detection, hinge-state classification, the active pipeline, and the VLM experiments.

3.1 Dataset

3.1.1 Dataset Design and Acquisition Protocol

For the purpose of this study, a dedicated image dataset was created, composed of five categories of handheld surgical instruments: needle holders, shears, tweezers, scalpels, and clamps. This selection follows the broad surgical instrument categorization described by Liehn and Schlautmann [LS22]. The physical instruments used in the dataset are listed in Appendix A.

The dataset was designed to perform controlled experiments on instrument recognition under variation in viewpoint, background, lighting, and hinge state. Three viewpoints were used: top-down (`TOP`), oblique (`OBL`), and close-up (`CLO`). Four background conditions were used: stainless steel tray (`TR`), green drape (`GR`), blue drape (`BL`), and white surface (`WH`). For hinged instruments, namely needle holders, shears, and clamps, both open (`OP`) and closed (`CL`) states were captured. Non-hinged instruments, namely tweezers and scalpels, were assigned the state label not applicable (`NA`).

The choice of including multiple viewpoints was made to evaluate how camera perspective affects recognition performance. Background variation was also included to test model robustness to changes in color, texture, and reflection. For hinged instruments, open and closed states were included to enable a separate hinge-state classification task. Open state images were captured with varying hinge angles, including small openings close to the closed state and unusually large openings, to increase variation within the open class.

The main dataset was collected under a controlled acquisition protocol shown in Table 1. For each instrument category, images were recorded across all combinations of viewing angles and background. Within each configuration, the instrument was repositioned and reoriented between each capture in order to introduce variation while preserving the controlled setup.

For each configuration, images were captured with balanced coverage across two different lighting conditions: artificial LED lighting and natural daylight. The artificial illumination was produced using a desk lamp, and the natural daylight condition used ambient lighting during daytime without any additional illumination. For each combination of instrument-viewpoint-background, approximately half of the images were acquired under artificial light, and half under natural daylight. Lighting was varied in order to introduce more diversity in illumination, but was not treated as

Table 1. Capture workflow followed during dataset acquisition. The procedure was repeated for each instrument category across all viewpoint and background combinations.

Step	Description
1	Select one instrument category and one physical instrument instance.
2	**for each** viewpoint $v \in \{\text{TOP}, \text{OBL}, \text{CLO}\}$ **do**
3	**for each** background $b \in \{\text{TR}, \text{GR}, \text{BL}, \text{WH}\}$ **do**
4	Place a single instrument within the image frame.
5a	**if** the instrument is hinged **then** acquire 21 open-state and 21 closed-state images.* Randomly vary the hinge angle for open-state captures.
5b	**else** acquire 42 images with the state label set to `NA`.*

* For each acquisition set, the instrument was repositioned and reoriented between captures, with approximately half of the images taken under artificial LED light and half under natural daylight.

a separate component in the dataset structure to keep dataset structure manageable. For hinged instruments, open and closed states were represented in equal proportion within the dataset.

To preserve traceability throughout preprocessing and evaluation, each image was saved using the following structured filename convention:

`[instrument]_[viewpoint]_[background]_[state]_[index].jpg`

For example, `CM_CLO_BL_OP_009.jpg` denotes the ninth image (`009`) of a clamp (`CM`) recorded from the close-up viewpoint (`CLO`) on the blue background (`BL`) in the open state (`OP`). Table 2 summarizes the meaning of each filename component.

Table 2. Filename attributes used in the dataset naming convention.

Attribute	Codes	Meaning
Instrument	NH, SH, TW, SC, CM	Needle holder, shears, tweezers, scalpel, clamp
Viewpoint	TOP, OBL, CLO	Top-down, oblique, close-up
Background	TR, GR, BL, WH	Stainless steel tray, green drape, blue drape, white surface
State	OP, CL, NA	Open, closed, not applicable / non-hinged
Index	001–042	Running image number within each setup

The final dataset (Figure 2) contains 2520 images distributed evenly across five surgical instrument categories. Images are evenly distributed across three viewpoints and four background conditions, with 504 images per instrument and 42 images per instrument-viewpoint-background combination, resulting in a total of 60 instrument-viewpoint-background combinations. Hinged instruments were further balanced equally across open and closed states. A complete breakdown of the full dataset is provided in Appendix B.1.

(a) NH, TOP, TR, OP **(b)** NH, OBL, BL, CL **(c)** CM, CLO, BL, CL

(d) TW, TOP, GR, NA **(e)** SH, OBL, WH, OP **(f)** SC, CLO, TR, NA

Figure 2. Representative samples from the full surgical instrument dataset showing variation in instrument category, viewpoint, background, hinge state, and illumination conditions. Abbreviations follow the dataset naming convention in Table 2.

3.1.2 Dataset Annotation

After data collection, images in the main dataset were manually annotated using the open-source labeling tool Label Studio [TMHL25], shown in Figure 3. Each image was assigned an instrument category label and a bounding box around the visible instrument.

To accelerate the annotation process, labeling was performed iteratively. The raw dataset was divided into four different batches (Table 3). As labeling each individual image manually is a time-consuming process, a first batch of 288 images was initially labeled manually. Once correctly labeled, this subset was used to train an early YOLO26n model. This model was subsequently used to predict the labels of the second batch of images. Although this early model was not yet highly accurate, its predictions provided useful initial bounding boxes and labels, which made the correction process faster than annotating each image from scratch.

After the second batch had been verified and corrected, its images and corresponding YOLO-format `.txt` annotation files were added to the `labeled_pool` folder together with the first batch. This enlarged labeled pool was then used to train a new detection model, which generated preliminary annotations for the third batch. The same procedure was repeated for all four batches. As the labeled pool increased in size, the annotation quality of the intermediate models improved. By the fourth batch, the annotation process mainly consisted of validating and correcting the annotations generated by the model rather than manually creating the bounding boxes and labels.

Once all four batches were added to the `labeled_pool` folder, final checks were performed. A

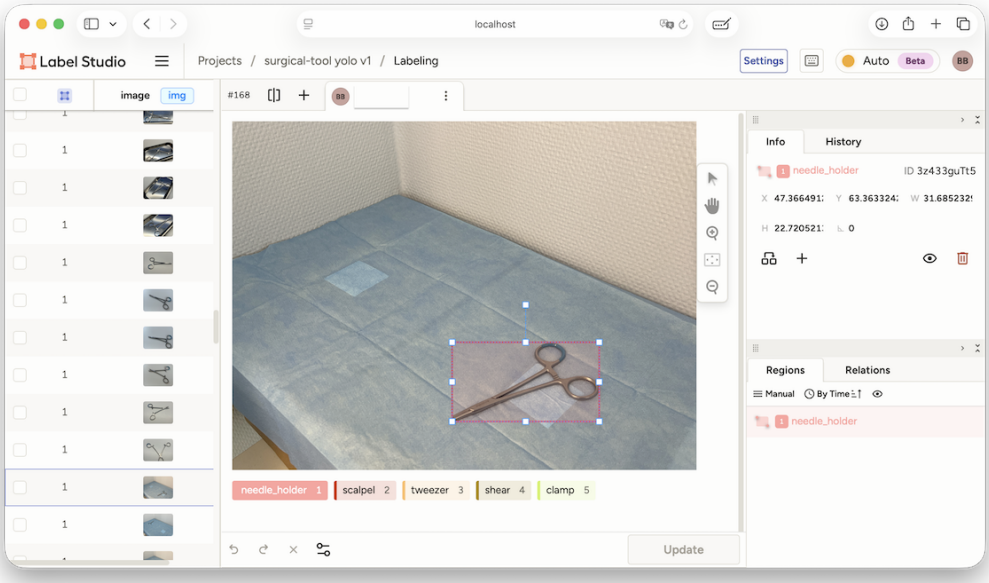

Figure 3. Label Studio interface used for dataset annotation. Each image was manually annotated with a bounding box and assigned to one of the predefined surgical instrument classes. In this example, a *needle_holder* instance is selected, together with its associated region metadata.

validation script was written to check the annotation consistency and integrity of the full dataset. The script checked that each image had a corresponding annotation file in YOLO format, that filenames all followed the expected structure with metadata, and that tool category, viewpoint, background, state, and index values were valid. It also verified that hinged and non-hinged instruments used appropriate state labels, that all expected index ranges were present, and that each annotation file contained exactly one valid bounding box with the class ID matching the instrument encoded in the file name. Finally, the script used SHA-256 hashes to identify any duplicate files with identical content. This validation step was used to ensure that the final labeled dataset was complete, consistent, and reliable before it was used for model training and evaluation.

Table 3. Batch sizes used during iterative dataset annotation.

Batch	**Number of images**
Batch 1	288
Batch 2	192
Batch 3	1008
Batch 4	1032
Total	2520

3.1.3 Supplementary Datasets

In addition to the main dataset, three additional datasets were created to run specific experiments for this study. These datasets were created for different purposes, including hinge-state classification, multi-view evaluation, and generalization testing. Together they allow to assess the proposed methods under a wider range of experiments than solely on the main dataset.

First, a cropped subset of the full dataset was created (Appendix B.2). This dataset consists of image crops centered on the instrument region and was used for the training and evaluation of hinge-state classification models. By removing the surrounding scene, and focusing on only the instrument, this dataset allowed the classification models to focus on the local visual features of the surgical instruments that are relevant to the open and closed hinge configurations. The dataset was used to train the vision-only hinge classifiers and to evaluate both the vision-only and vision-language models on the hinge-state recognition task.

Second, an extended dataset was constructed in which images of the same scene were linked across viewpoints (Appendix B.3). In this dataset, corresponding top-down, oblique, and close-up images show the same instrument configuration. These scene-linked images were necessary for experiments that required multi-view inference, since it allows for direct comparison of the predictions from multiple viewpoints. It also allows for the fusion of predictions when considered at the scene level, and not solely the image level. The extended dataset was therefore used to evaluate whether combining multiple views improves recognition performance over single-view information.

Lastly, a separate generalization dataset (Appendix B.4) was created to evaluate how well the trained models transfer to unseen instrument instances. This dataset contains instruments that belong to the same broad categories as those in the full dataset, but uses different instances of physical objects from the ones used in training. As a result, it provides a more challenging test of generalization. This dataset was used to evaluate whether the models learned features relevant to category and state that can be applied to new instruments of the same class.

3.2 Vision-Only Models

This section describes the vision-only models trained for the two recognition tasks in this thesis: instrument category detection and hinge-state classification. It introduces the selected detection and classification models, followed by the training setup used.

3.2.1 YOLO26 Object Detection Models

For the instrument detection task, two object detection models from the YOLO26 family were trained to localize the instrument in each image, and to assign it to one of the five predefined instrument labels. This detection stage is the first component of the overall recognition pipeline as its predicted bounding box is also used to generate the crop for hinge-state classification. Two detection model variants were used: YOLO26n and YOLO26s. These models were chosen to compare two models of the same family with different architecture sizes. Both models were initialized from pretrained weights and trained on the same detection dataset derived from the full labeled set.

Table 4. Overview of the vision-only object detection models evaluated in this study. Both models were initialized from COCO-pretrained weights and trained using an input size of 640 × 640. Parameter counts are reported in millions.

Model	Variant	Parameters
YOLO26n [Ult25c, Ult25b]	Nano	2.51 M
YOLO26s [Ult25c, Ult25b]	Small	9.95 M

3.2.2 Hinge-State Classification Models

In addition to object detection, recognition of instrument state for hinged surgical tools was evaluated. This task is formulated as a binary image classification problem in which the model predicts whether a detected instrument is in an open or closed configuration. Compared with category recognition, this task requires attention to more geometric visual cues, such as hinge angle, jaw separation, and tip shape.

The four classification models used were EfficientNet-B0 [TL19], ResNet-18 [HZRS16], and two YOLO classification models: YOLO26n-cls and YOLO26s-cls [Ult25a]. These models were selected to compare different vision model families under the same binary classification setup. ResNet-18 was included as a widely used convolutional baseline model, while EfficientNet-B0 was included as a compact architecture designed for efficient scaling [TL19]. The two YOLO classification variants were included to assess whether models from the same family as the detector also perform well on the hinge-state task.

Table 5. Overview of the vision-only hinge-state classification models evaluated in this study. All models were trained for 40 epochs using an input size of 224 × 224 and ImageNet-pretrained weights for initialization. Parameter counts are reported in millions.

Model	Family	Parameters
EfficientNet-B0 [TL19]	EfficientNet	5.29 M
ResNet-18 [HZRS16]	ResNet	11.69 M
YOLO26n-cls [Ult25c, Ult25a]	YOLO classification	2.8 M
YOLO26s-cls [Ult25c, Ult25a]	YOLO classification	6.7 M

3.2.3 Training Setup

All trainable vision models were evaluated using fixed train, validation, and test splits. The models were not trained from scratch, and instead transfer learning was used: each model was initialized from publicly available pretrained weights and fine-tuned on the custom surgical instrument datasets collected in this thesis.

For the object detection task, the full labeled dataset was split according to an 80:10:10 train-validation-test ratio. Before splitting, the dataset was checked to ensure that each image had a corresponding annotation file in YOLO format and that all images had a matching label. The

image-label pairs were then randomly shuffled and copied into the corresponding train, validation, and test folders. The two detection models, YOLO26n and YOLO26s, were initialized from the official pretrained detection checkpoints `yolo26n.pt` and `yolo26s.pt` [Ult25b]. These checkpoints were pretrained on the COCO object detection dataset [LMB+14]. The detection heads were then adapted to the five surgical instrument categories used in this thesis, and the models were fine-tuned on the collected full dataset for 80 epochs using an input size of 640×640 and a batch size of 16.

For the hinge-state classification task, the cropped hinged-instrument dataset was derived from the already split object detection dataset rather than split independently. The four classification models were initialized from image classification weights pretrained on ImageNet, the large-scale image classification dataset introduced by Deng et al. [DDS+09]. EfficientNet-B0 and ResNet-18 used ImageNet-1K pretrained weights provided through TorchVision [Tor26], while YOLO26n-cls and YOLO26s-cls used the Ultralytics classification checkpoints pretrained on ImageNet [Ult25a]. The classification heads were adapted to the binary open/closed hinge-state task, and all four classifiers were fine-tuned on the cropped hinge dataset for 40 epochs using an input size of 224×224 and a batch size of 32.

3.3 Active Vision Pipeline

The active vision pipeline is described in three stages. First, Section 3.3.1 presents the single image inference pipeline, which combines instrument detection with hinge-state classification. Second, Section 3.3.2 describes how predictions from multiple viewpoints are fused into a single scene-level output. Finally, Section 3.3.3 introduces the threshold-based active acquisition policy, which determines whether additional viewpoints are needed before returning a final prediction.

3.3.1 Unified Detection and State Classification Pipeline

The proposed system is based on a unified inference pipeline that combines instrument detection with hinge-state classification (Figure 4). This pipeline was implemented both in an interactive application, which operates on live camera input, and in a controlled evaluation setup used for more systematic and formal multi-view experiments with still images.

For each image input, the detector predicts bounding boxes, instrument categories, and detection confidence scores. As each image contains a single main instrument, only the highest confidence detection is retained. If the predicted category belongs to a hinged instrument class, the corresponding bounding box is used to crop the detected image region. This crop is then passed to a binary hinge-state classifier which predicts whether the instrument is in the closed or open configuration. For non-hinged instruments, no additional state prediction is performed. The final pipeline output is therefore the predicted instrument category and bounding box, together with a hinge-state prediction and classification confidence when applicable.

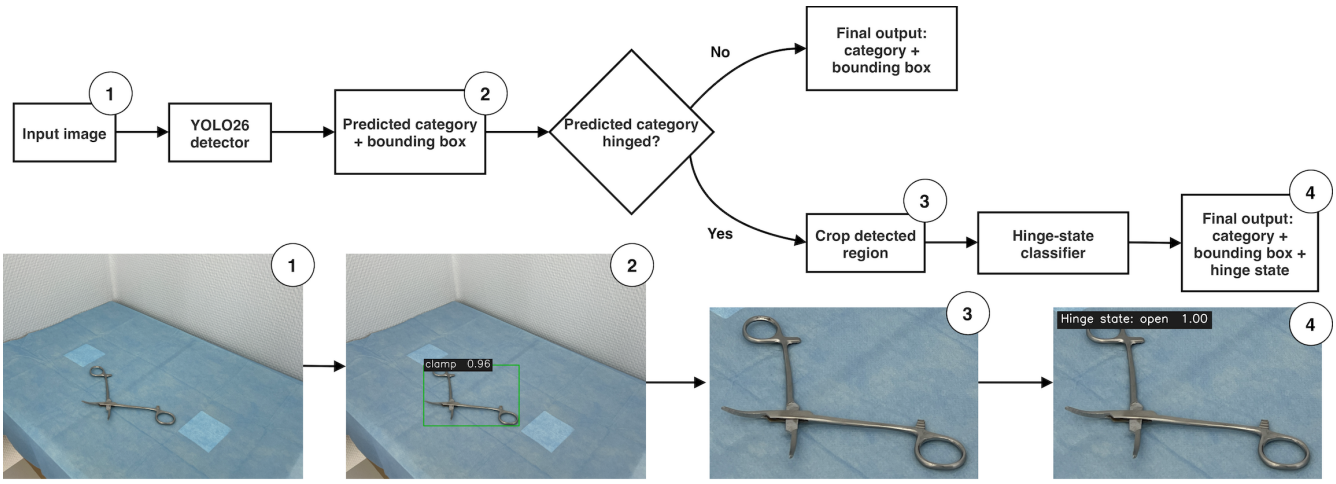

Figure 4. Unified inference pipeline for instrument detection and hinge-state classification. The upper row shows the decision flow of the pipeline, while the lower row illustrates the corresponding inference steps on an example clamp image. The YOLO26 detector predicts the category and bounding box. If the predicted category is hinged, the detected region is cropped and passed to the hinge-state classifier to produce the final category, localization, and hinge-state output.

3.3.2 Multi-view Fusion

When one or more views of the same scene are available, their predictions are combined through multi-view fusion to produce a single scene-level output rather than treating each viewpoint independently. Each scene can be observed from the three predefined viewpoints introduced in Section 3.1: top-down (`TOP`), oblique (`OBL`), and close-up (`CLO`).

The set of acquired views for a given scene is denoted V, which may contain a single view, all three views, or a subset selected by the active acquisition strategy described in Section 3.3.3. For each view $v \in V$, the unified pipeline is applied independently. Since each image contains one main instrument, only the highest confidence detection is retained, yielding a predicted instrument category, and a detector confidence score. Multi-view fusion then aggregates these view-level predictions into a single scene-level output.

Instrument category fusion For category recognition, fusion is performed by summing the detector confidence scores across the acquired views for each predicted class. This summation fusion strategy is consistent with classical work on classifier combination, where the sum rule has been shown to be a simple and robust way to combine multiple classifier outputs [KHDM98]. Letting F_k denote the fused score for instrument class k,

$$F_k = \sum_{v \in V_k} c_v,$$

where V_k is the set of views whose top detection predicted class k, and c_v is the corresponding confidence score in view v. The final tool prediction is taken as the class with the highest fused score.

Hinge-state fusion For scenes with hinged instruments, fusion is performed by summing the predicted probabilities for the closed and open states across views:

$$S_{\mathrm{CL}} = \sum_{v \in V} p_v(\mathrm{CL}), \qquad S_{\mathrm{OP}} = \sum_{v \in V} p_v(\mathrm{OP}),$$

where $p_v(\mathrm{CL})$ and $p_v(\mathrm{OP})$ are the predicted probabilities of the closed and open states in view v, respectively. The fused hinge-state prediction is taken as the state with the higher aggregated score.

Fusion confidence scores In addition to the fused predictions, confidence scores are defined for both tasks and used by the active acquisition policy to determine whether an additional viewpoint is needed.

For tool recognition, confidence is measured as the gap between the highest and second-highest fused scores, normalized by the number of acquired views:

$$c_{\mathrm{tool}} = \frac{F_{(1)} - F_{(2)}}{|V|},$$

where $F_{(1)}$ and $F_{(2)}$ are the highest and second-highest fused tool scores, and $|V|$ is the number of acquired views. Larger values indicate a larger separation from other tool classes, and therefore a higher confidence.

For hinge-state classification, confidence is defined as

$$c_{\mathrm{state}} = \frac{\max(S_{\mathrm{CL}}, S_{\mathrm{OP}})}{S_{\mathrm{CL}} + S_{\mathrm{OP}}},$$

with larger values indicating that one state is clearly favored over the other after aggregating predictions across views.

3.3.3 Threshold-Based Active View Acquisition

Within the controlled evaluation, a threshold-based active acquisition policy is defined to determine whether additional viewpoints are required. The available viewpoints are considered in a fixed order:

$$\mathrm{TOP} \rightarrow \mathrm{OBL} \rightarrow \mathrm{CLO}.$$

The top-down view was therefore always used as the initial observation, while the oblique and close-up views were acquired only when the current fused prediction did not satisfy the stopping criteria.

The active policy, summarized in Figure 5 and Algorithm 1, was defined using confidence thresholds for tool and state prediction to determine whether additional viewpoints were required. If the predicted tool was non-hinged, the process stopped once the fused tool confidence exceeded the tool threshold τ_{tool}. If the predicted tool was hinged, both the fused tool confidence and the fused state confidence had to exceed their respective thresholds τ_{tool} and τ_{state}. If these conditions were

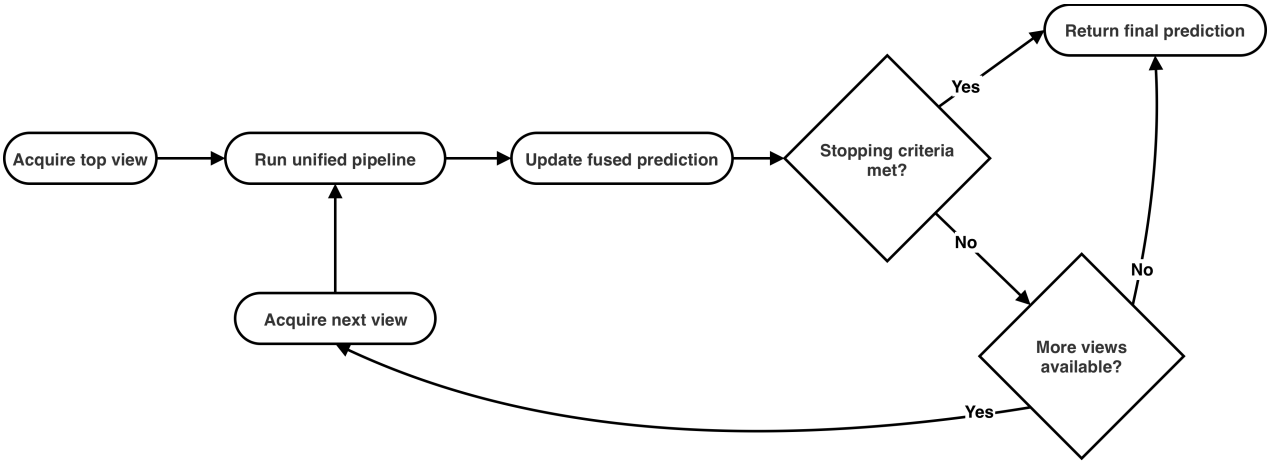

Figure 5. Threshold-based active view acquisition strategy used for multi-view inference. The system starts from the top-down view and applies the unified pipeline to each acquired image. After each view, the fused prediction is updated and the stopping criteria are evaluated. If the prediction is sufficiently confident, the current fused prediction is returned; otherwise, the next available viewpoint is acquired. If no further viewpoints are available, the final fused prediction is returned.

Algorithm 1: Threshold-based active multi-view acquisition

Input: tool threshold τ_{tool}, state threshold τ_{state}
foreach $v \in \{\text{TOP}, \text{OBL}, \text{CLO}\}$ **do**
  Acquire view v;
  Update fused tool prediction and fused tool confidence c_{tool};
  if *predicted tool is hinged* **then**
    Update fused state prediction and fused state confidence c_{state};
  if *predicted tool is non-hinged* ***and*** $c_{\text{tool}} \geq \tau_{tool}$ **then**
    return fused tool prediction;
  if *predicted tool is hinged* ***and*** $c_{\text{tool}} \geq \tau_{tool}$ ***and*** $c_{\text{state}} \geq \tau_{state}$ **then**
    return fused tool and state prediction;
return final fused prediction after all three views;

not met, the next predefined viewpoint was acquired and added to the fused prediction. If all three views had been used, the final fused prediction was returned regardless of confidence.

In addition to the active policy, a non-active multi-view variant was defined in which all three available views of a scene were systematically fused. The evaluation procedure used to assess the active acquisition strategy is later described in Section 3.5.3.

3.4 Vision-Language Models

In addition to trained vision-only models, VLMs were also evaluated on instrument category recognition and hinge-state classification tasks. This comparison was included in order to assess whether general-purpose VLMs can perform the same recognition tasks without training specifically for the task.

3.4.1 Evaluated VLMs

Several VLMs were selected to cover multiple model families and parameter scales while remaining computationally feasible to evaluate locally. The evaluated models include compact models, such as Gemma 3n E4B and Ministral 3 3B, as well as larger models such as Gemma 3 12B and Qwen3-VL 8B. The selected models all support image input, but differ in scale, latency, and family architecture, which allows comparison of the performance and efficiency of these different models.

Table 6. Overview of the VLMs evaluated in this study. All models were run as locally installed LM Studio checkpoints using 4-bit quantized variants and support image input.

Model	Parameters	Architecture	Format
Gemma 3 12B [Goo25a]	12B	Gemma 3	SafeTensors
Gemma 3n E4B [Goo25b]	4B	Gemma 3n	SafeTensors
Gemma 4 E4B [Goo26]	7.5B	Gemma 4	GGUF
Ministral 3 3B [Mis25]	3B	Ministral 3	GGUF
Qwen3-VL 4B [Qwe25a]	4B	Qwen3-VL	SafeTensors
Qwen3-VL 8B [Qwe25b]	8B	Qwen3-VL	SafeTensors

The VLMs were evaluated on two tasks: instrument category recognition in Section 4.4.1 and hinge-state classification in Section 4.4.2.

3.4.2 Prompting Strategy and Output Normalization

The VLMs were evaluated in a zero-shot setting using predefined prompts on the closed set classification setup. In all runs, each request consisted of a single image with a short text instruction, and the models were required to return exactly one label from a predefined set of allowed categories. To ensure consistency across models, temperature was fixed to 0.0.

For the instrument category recognition task, two prompts were considered. The baseline prompt instructed each model to identify the instrument and return exactly one label from the five predefined categories. Additionally, an optimized prompt was evaluated. This prompt provided a short visual description of each category, and included more explicit guidance on how to classify the instruments. The purpose of this optimized prompt was to investigate whether the quality of a prompt had a significant impact on recognition of visually similar instruments.

For the hinge-state classification task, the prompting setup chosen was simpler. Each input consisted of a cropped image of a hinged surgical instrument, and the model was instructed to return exactly one of the two labels: `open` or `closed`.

Because VLM outputs may vary in wording or formatting, a rule-based output normalization step was also applied before evaluation of the results. For the instrument category task, responses were converted to lowercase and mapped to the predefined labels using a small set of similar words, which allowed variation such as singular and plural forms (e.g. `tweezer` and `tweezers`). For the hinge-state task, the normalization step was stricter, and output was converted to lowercase, and also stripped of whitespace. Therefore, only exact matches to `open` and `closed` were accepted. Outputs that did not match to a valid label were marked as invalid predictions.

3.5 Evaluation Protocol

Evaluation of the models was done in stages. First, the object detection models and hinge-state classification models were evaluated separately. Second, the proposed unified pipeline was evaluated with multi-view fusion and threshold-based active view acquisition. Finally, the VLMs were evaluated as benchmarks for comparison with the trained vision-only models.

3.5.1 Object Detection Evaluation

The object detection models were evaluated on the test split of the full dataset. As each test image contains exactly one annotated surgical instrument, the task was to both localize the object in the image and assign it to the correct instrument category. Performance was measured using precision, recall, mean Average Precision at an IoU threshold of 0.50 (mAP@50), and mean Average Precision averaged over IoU thresholds from 0.50 to 0.95 (mAP@50:95). These metrics were chosen because they capture both classification performance and localization quality, and are widely used for the evaluation of object detection models.

In addition to these overall metrics, precision by class, recall, F1-score, mAP@50, and mAP@50:95 were reported for the five instrument categories. Confusion matrices were also included to provide further insight into which categories were most often confused. Lastly, to complement recognition performance, inference speed was measured as a mean latency per image on the local setup using Apple MPS (Metal Performance Shaders). This metric was included because the intended application of the model is a practical vision system, in which latency or response speed is as important as recognition precision.

3.5.2 Hinge-State Classification Evaluation

The hinge-state classification models were evaluated as a binary image classification task on cropped images of hinged surgical instruments: clamps, needle holders, and shears. Each crop was labeled as either `open` or `closed`. By using cropped instrument regions instead of full-scene images, the evaluation focused specifically on the local visual cues relevant to hinge-state recognition, such as jaw separation, hinge angle, and the geometry of the tips.

The evaluation was performed on the test split of the cropped hinge-state dataset, which contained 157 images in total, consisting of 78 open and 79 closed samples. The main evaluation metric was overall classification accuracy on the full test set. In addition, precision, recall, and F1-score were reported for the open and closed classes. Confusion matrices were also included to provide further insight into the types of errors made by the models.

3.5.3 Unified Pipeline Evaluation

The unified pipeline was evaluated on the same-scene extended dataset described in Section 3.1.3, in which each instrument was captured from three viewpoints: top-down, oblique, and close-up, without moving the object between captures. This setup enables evaluation per scene while keeping the instrument placement and configurations fixed across views.

Three inference settings were compared: single-view inference, full multi-view fusion, and threshold-based active view acquisition. In the single-view setting, each viewpoint was evaluated independently. In the full fusion setting, all three views of the same scene were combined using the multi-view fusion procedure described in Section 3.3.2. In the active setting, the system started from the top-down view and acquired the oblique and close-up views only when the confidence criteria defined in Section 3.3.3 and Algorithm 1 were not satisfied.

To evaluate the active acquisition method, a sweep over stopping threshold was performed. Tool confidence thresholds ranged from 0.30 to 0.95 in steps of 0.05, while state confidence thresholds ranged from 0.50 to 0.95 in steps of 0.05, resulting in 140 threshold pairs for each detector-classifier model combination. For each threshold pair, the active policy was simulated using the fixed order of viewpoints.

Performance was then evaluated at the scene level. Tool accuracy was computed over all scenes by comparing the final predicted category with the ground truth category. For hinged instruments, two additional metrics were reported. First, conditional state accuracy was computed. This represents hinged scenes in which the tool was correctly recognized. This shows how well the system was able to predict the hinge state once the tool has been correctly identified. Second, end-to-end state accuracy was computed over all hinged scenes, which requires both the tool category and the hinge state to be correct. Finally, the mean number of acquired views was also reported as a measure of how efficient the active acquisition strategy is compared to the other methods.

3.5.4 Vision-Language Models Evaluation

The VLMs were evaluated as zero-shot classifiers on the test splits that correspond to each task. Instrument category recognition was evaluated on the test split of the full dataset, while hinge-state classification was evaluated on the test split of the cropped hinged instrument dataset. In both cases, the normalized model output was used to compare it to the ground truth label. For instrument category recognition, both the baseline and optimized prompts described in Section 3.4 were evaluated. For hinge-state classification, only one prompt was used.

Performance was summarized using four metrics: strict accuracy, valid prediction rate, valid-only accuracy, and mean inference latency. Strict accuracy was computed by counting invalid outputs as incorrect predictions, while valid prediction rates measured the proportion of outputs that could be mapped to one of the allowed labels. Valid-only accuracy was calculated only over the valid predictions, which separated recognition difference from failures to follow prompts. Mean inference latency was reported to compare the computational efficiency of each model.

4 Results

This section presents the results in the same order as the methodology. First, the individual vision-only models are evaluated. The integrated active multi-view pipeline is then assessed, followed by the VLM results. Finally, additional analyses are presented to examine interpretability, generalization, and edge cases.

4.1 Object Detection Results

4.1.1 Full-Test Performance

The two evaluated object detectors, YOLO26n and YOLO26s, both achieved strong performance on the main surgical instrument dataset test split. As shown in Table 7, both models reached precision and recall above 0.98, as well as mAP@50 values above 0.99. This indicates that both models can localize the object and predict the category with high accuracy.

Table 7. Comparison of YOLO26n and YOLO26s on the surgical instrument test split. Inference time per image was measured on the local test setup using Apple MPS.

Model	Params	Precision	Recall	mAP@50	mAP@50:95	Time (ms)
YOLO26n	2.51M	0.985	0.984	0.995	0.982	51.32
YOLO26s	9.95M	0.989	0.985	0.994	0.981	58.33

The confusion matrices in Figure 6 confirm that the number of classification errors was small. Most samples were assigned to the correct class, and missed detections were rare. There were however some cases of false positives, where the model predicted instruments to exist in background regions with no actual instrument present.

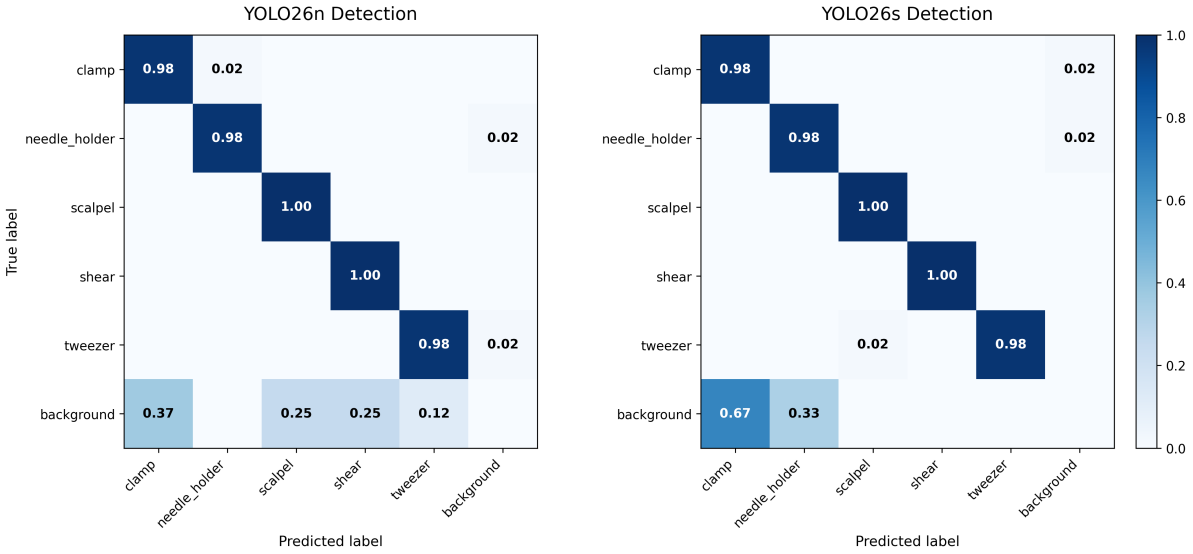

Figure 6. Normalized confusion matrices for the two object detection models on the test set (n = 252: 56 clamp, 53 needle holder, 41 scalpel, 48 shear, and 54 tweezer instances). Rows indicate true labels and columns indicate predicted labels. The background row indicates false positive detections and the background column indicates missed detections. YOLO26n achieved the highest mAP@50–95 at 98.2%.

The performance by class in Table 8 shows that performance was consistently strong across all instrument types for both detectors. Scalpels and shears achieved perfect recall on the test set, and needle holders also achieved a very high score. The tweezers and clamps were also very high, albeit lower than the other categories. These results suggest that the confusions were limited when detecting surgical instruments in this setup.

Table 8. Per-class test results for YOLO26n and YOLO26s. Recall and mAP@50:95 are shown to highlight class-wise robustness differences.

Class	Recall (n)	Recall (s)	mAP@50:95 (n)	mAP@50:95 (s)
Clamp	0.963	0.977	0.986	0.985
Needle holder	0.981	0.981	0.991	0.989
Scalpel	1.000	1.000	0.972	0.974
Shear	1.000	1.000	0.993	0.990
Tweezer	0.977	0.969	0.966	0.965

4.1.2 Performance by Viewpoint

In order to investigate the importance of viewpoint in object detection, the test results were analyzed separately for each view. The results by viewpoint are reported in Table 9 and visualized in Figure 7. Accuracy was high across all viewpoints, which indicates that the models were robust to changes in viewing angles.

Performance by viewpoint shows that both YOLO object detection models are highly accurate

Table 9. Object detection performance by viewpoint for YOLO26n and YOLO26s on the test split.

Model	Viewpoint	n	Precision	Recall	mAP@50	mAP@50:95
YOLO26n	TOP	97	0.940	0.991	0.993	0.973
YOLO26n	OBL	78	0.989	0.993	0.995	0.978
YOLO26n	CLO	77	0.989	0.998	0.995	0.992
YOLO26s	TOP	97	0.984	0.997	0.995	0.981
YOLO26s	OBL	78	0.961	0.989	0.990	0.971
YOLO26s	CLO	77	0.993	0.990	0.995	0.990

across all three camera perspectives on the test split. For both YOLO26n and YOLO26s, the close-up view produced the strongest results overall, achieving the highest or near-highest scores across all metrics. For YOLO26n, the oblique view performed better than the top-down view, whereas for the larger YOLO26s model, the top-down view performed better than the oblique view.

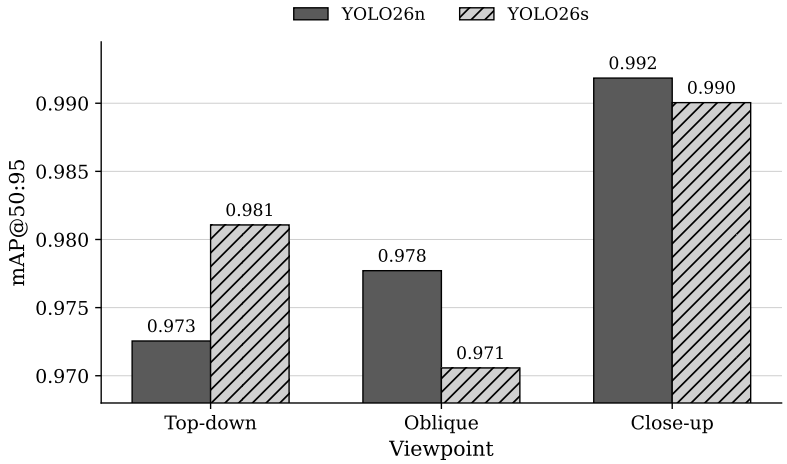

Figure 7. Object detection performance by viewpoint for YOLO26n and YOLO26s, measured using mAP@50:95 on the test split. Close-up views yielded the strongest overall performance for both models, while oblique and top views performed second respectively for YOLO26n and YOLO26s.

Overall, the results by viewpoint for object detection show that while accuracy was high for all images, close-up images provide the best conditions for accurate recognition of the instrument.

4.2 Hinge-State Classification Results

4.2.1 Full-Test Performance

The hinge state classification models were tested on the test split of the dataset. The test set was obtained by doing an 80:10:10 split on the cropped instruments. The test set had a size of 157 images, with 79 closed and 78 open instruments in three different categories of instruments: needle holder, clamp, and shear. Figure 8 shows the confusion matrices of the four classification models,

using the same training, validation, and testing sets.

Table 10. Comparison of hinge-state classification models on the test set ($N = 157$).

Model	Accuracy (%)	Recall (open) (%)	Recall (closed) (%)
EfficientNet-B0	98.09	96.15	100.00
YOLO26n-cls	97.45	96.15	98.73
ResNet-18	97.45	94.87	100.00
YOLO26s-cls	96.82	94.87	98.73

EfficientNet-B0 performed the best on the hinge-state classification test set, reaching 98.09% accuracy and a macro F1-score of 98.09%. YOLO26n-cls and ResNet-18 followed closely with identical accuracies of 97.45%, while YOLO26s-cls obtained 96.82%. Across all evaluated models, recall was consistently higher for the closed class than for the open class, suggesting that open configurations showed more variability and were therefore more difficult to classify reliably.

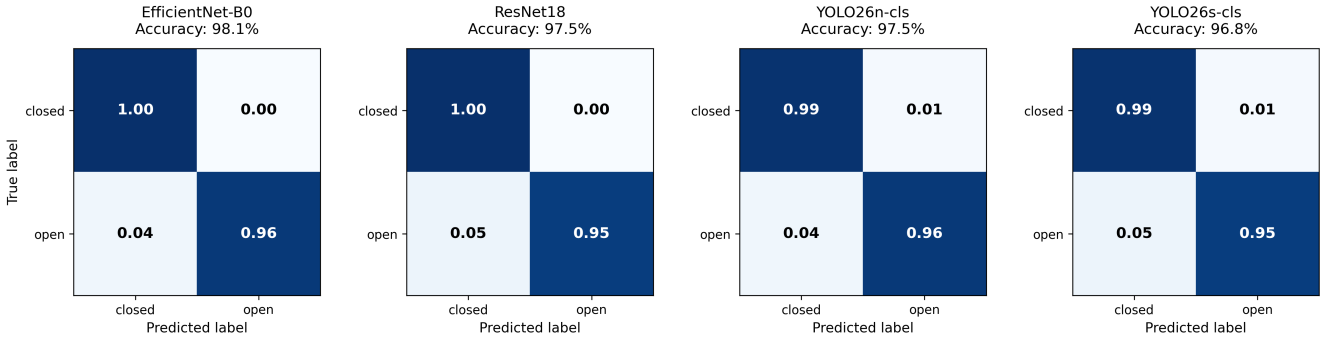

Figure 8. Confusion matrices for the four hinge-state classification models on the test set (n = 157, 79 closed and 78 open). Rows indicate true labels and columns indicate predicted labels. EfficientNet-B0 achieved the highest accuracy at 98.1%.

EfficientNet-B0 achieved the highest accuracy at 98.1%, followed by ResNet-18, YOLO26n-cls, and finally YOLO26s-cls.

4.2.2 Performance by Viewpoint

Performance by viewpoint showed that the hinge-state classification task followed a different trend from the object detection task. For EfficientNet-B0, ResNet-18, and YOLO26n-cls, the highest accuracy was obtained on the top-down view, where all three models reached 100.0% accuracy. In contrast, YOLO26s-cls performed best on the oblique view with an accuracy of 97.8%. Across all models, close-up views yielded the weakest results.

This suggests that hinge-state recognition depends less on fine close-up detail than on a clear view of the overall geometric configuration of the instrument. A top-down perspective may make the separation between the two halves of the hinged instruments and their angle easier to observe, which is relevant information to determine whether the instrument is in an open or closed state.

Table 11. Hinge-state classification accuracy (%) by viewpoint on the cropped test set. The number of test images per viewpoint was 64 for top-down, 46 for oblique, and 47 for close-up.

Model	Top-down	Oblique	Close-up
EfficientNet-B0	**100.00**	97.83	95.74
ResNet-18	**100.00**	95.65	95.74
YOLO26n-cls	**100.00**	97.83	93.62
YOLO26s-cls	96.88	**97.83**	95.74

The oblique view may provide useful depth information, but its advantage appears to depend on the model architecture. Overall, these results indicate that viewpoint has a measurable effect on hinge-state classification.

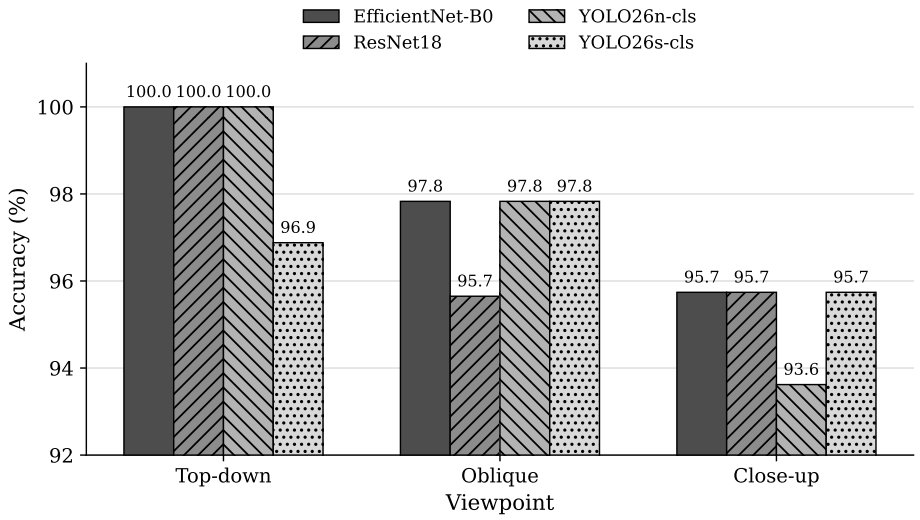

Figure 9. Hinge-state classification accuracy by viewpoint for the four evaluated classifiers. Top-down views yielded the strongest performance for EfficientNet-B0, ResNet-18, and YOLO26n-cls, while YOLO26s-cls performed best on the oblique view.

It should also be noted that each viewpoint subset for the classification task was smaller than for the detection task, and the number of image samples per view is not perfectly balanced.

4.3 Active Multi-View Pipeline Results

This experiment evaluated whether recognition of the same scenes could be improved by acquiring additional viewpoints only when the current fused prediction was not confident enough. The analysis was performed on the extended same-scene dataset (Appendix B.3) which contained 100 challenging scenes, each captured from three linked viewpoints. It is important to note that these scenes contain partial occlusions, lowering the accuracy of the models compared to the baseline dataset used in previous evaluations. For each detector-classifier combination, three settings were compared: top view only, three views full fusion, and a threshold-based active acquisition policy.

Unlike the cropped hinge-state evaluation in Table 11, this experiment evaluates the complete detection and classification pipeline at scene level. Therefore, the reported end-to-end hinge-state accuracy requires both the instrument category and the hinge state to be correct, and can be reduced by detector errors, incorrect crops, category confusions, occlusions, or hinge-state classification errors.

Table 12 summarizes the scene-level end-to-end hinge-state accuracy for these three settings. For all evaluated combinations, the active strategy outperformed the top view only baseline. Compared to the fusion of all three views, the active policy produced higher end-to-end accuracy in three of the eight detector-classifier combinations, matched it in one combination, and performed lower in the remaining four combinations.

The strongest active-policy result was obtained with the combination of the YOLO26n detector and the EfficientNet-B0 hinge-state classifier. For this pipeline, top-view-only inference achieved an end-to-end hinge-state accuracy of 0.550, full fusion of all three views increased this to 0.700, and the best active policy further improved it to 0.733 while using 2.50 views on average.

Table 12. Comparison of scene-level end-to-end hinge-state accuracy for top-view-only inference, full fusion of three viewpoints, and the best observed active threshold-based policy for each detector-classifier combination. The active-policy result corresponds to the best threshold pair found in the threshold sweep. The strongest active result was obtained by YOLO26n + EfficientNet-B0, which achieved 0.733 end-to-end accuracy while using 2.50 views on average. The same end-to-end accuracy was also obtained by the full-fusion YOLO26s + EfficientNet-B0 configuration.

Detector	State classifier	TOP-only E2E	Fused E2E	Active E2E	Mean views used
YOLO26n	EfficientNet-B0	0.550	0.700	**0.733**	2.50
YOLO26n	ResNet-18	0.500	0.700	0.717	2.50
YOLO26n	YOLO26n-cls	0.450	0.683	0.683	2.52
YOLO26n	YOLO26s-cls	0.483	0.667	0.717	2.50
YOLO26s	EfficientNet-B0	0.667	**0.733**	0.700	1.32
YOLO26s	ResNet-18	0.617	0.717	0.667	1.48
YOLO26s	YOLO26n-cls	0.600	0.717	0.700	2.41
YOLO26s	YOLO26s-cls	0.633	0.700	0.667	1.67

For the YOLO26s detector, the best full-fusion result was obtained with EfficientNet-B0, reaching an end-to-end hinge-state accuracy of 0.733 compared with 0.667 for top-view-only inference. The corresponding active policy achieved 0.700 while using only 1.32 views on average.

Overall these results show that information from additional views was beneficial at the scene level, but always fusing all three views was not consistently optimal. Selectively requesting additional viewpoints provided an accuracy-efficiency trade-off, improving over the initial top-view baseline while often using fewer views than full fusion.

4.4 Vision-Language Model Results

The VLM experiments were included to assess whether general-purpose multimodal models could provide comparable performance on zero-shot configurations of instrument category recognition

and hinge-state classification tasks. Overall, the results show that the VLMs were able to capture some distinctions between instruments, but remained clearly below the performance of specialized vision-only models, especially on the hinge classification task.

4.4.1 Tool Recognition with VLMs

Table 13 shows that the Qwen3-VL models achieved the strongest overall performance on the tool category recognition task. The best result with the base prompt was obtained by Qwen3-VL-8B with an accuracy of 72.2%, and the same model also achieved the best result with the optimized prompt at 71.0%. Qwen3-VL-4B followed closely under the optimized prompt with 69.8%. Gemma-3-12B performed moderately well, with an accuracy of 61.1% with the optimized prompt, whereas the other three models remained below 40% accuracy.

Table 13. Comparison of VLM performance on the surgical tool category recognition test set ($n = 252$) under the base prompt (BP) and optimized prompt (OP). Accuracy is reported over all test images. Valid rate indicates the proportion of outputs matching one of the allowed class labels. Latency denotes mean response time per image in seconds.

Model	**BP Acc.** (%)	**OP Acc.** (%)	**Δ Acc.** (pp)	**BP Valid** (%)	**OP Valid** (%)	**BP Lat.** (s)	**OP Lat.** (s)
Gemma-3-12B	50.0	61.1	+11.1	100.0	100.0	9.54	12.96
Gemma-3n-E4B	33.3	38.5	+5.2	79.0	100.0	2.50	3.28
Gemma-4-E4B	28.6	20.2	-8.3	81.3	57.9	3.71	4.62
Ministral-3-3B	27.0	33.7	+6.7	98.0	100.0	7.99	6.81
Qwen3-VL-4B	62.7	69.8	+7.1	100.0	100.0	26.68	28.43
Qwen3-VL-8B	72.2	71.0	-1.2	100.0	100.0	39.21	44.12

Prompt optimization improved performance for most models, with the exception of two (Figure 10). The largest accuracy gain was observed for Gemma-3-12B, which improved by 11.1 percentage points. Qwen3-VL-4B showed substantial improvement, while Gemma-3n-E4B and Ministral-3-3B showed smaller gains. In contrast, Qwen3-VL-8B decreased slightly under the optimized prompt, and Gemma-4-E4B significantly deteriorated in both accuracy and valid prediction rate. This suggests that prompt adaptation had a measurable effect, but that its benefit depended strongly on the model family.

A clear pattern in classes was observed in the stronger models. Visually distinct instruments such as tweezers, scalpels, and shears were recognized more reliably, while needle holders and clamps were more difficult to identify. The most common confusions occurred between needle holders and clamps, and to a lesser extent between clamps and shears. This suggests that VLMs were reasonably capable at categorizing instruments with large visual differences, but struggled with finer-grained details such as similar hinged tools.

Viewpoint and background also influenced the performance of VLMs. When averaged across different models close-up views yielded the strongest category recognition results, whereas top-down views

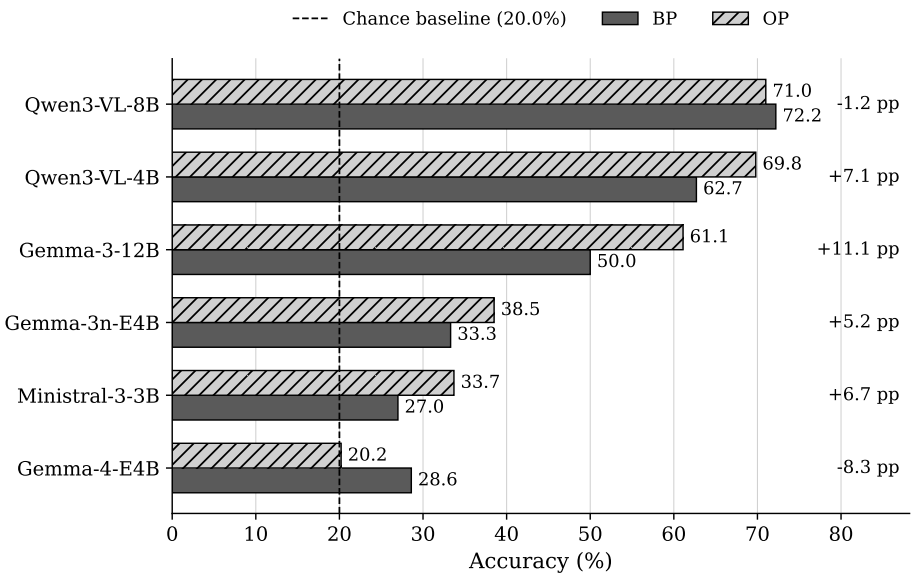

Figure 10. Comparison of VLM instrument category recognition accuracy under the base prompt (BP) and optimized prompt (OP). The dashed vertical line indicates the five-class chance baseline of 20%. Values on the right show the change in accuracy from BP to OP in percentage points.

were generally most difficult. The reflective tray background also tended to reduce accuracy when compared to other less reflective backgrounds.

Although the best VLMs achieved moderate performance, they remain clearly inferior to vision-only detectors, which reached near perfect recognition on the same test split. Furthermore, VLMs were substantially slower, with a mean latency ranging from a few seconds per image to more than 40 seconds for Qwen3-VL-8B. Overall, the VLMs provided a useful comparative baseline, but did not match the accuracy or speed of vision-only models.

4.4.2 Hinge-State Classification with VLMs

The hinge-state classification results were considerably weaker than the tool category recognition. The best-performing model was Qwen3-VL-8B, which achieved a strict accuracy of 62.4%. Gemma-3-12B followed with 55.4%, while Qwen3-VL-4B and Ministral-3-3B remained close to the level of a coin toss. Gemma-3n-E4B fell below chance level, and Gemma-4-E4B performed particularly poorly under the strict metric because many of its outputs did not conform to the required label format. Full results are reported in Appendix D.5.

These results show that hinge-state recognition was a difficult task for the evaluated VLM models. While the strongest performed above 60% strict accuracy, none of the VLMs achieved performance comparable to the dedicated classifiers. The vision-only classifiers all reached accuracies above 96%, showing that VLMs are not competitive at this specific task.

Another failure case was response bias. Several models showed a strong tendency to favor one class, regardless of the state of the instrument on the image. This behavior was especially visible for Ministral-3-3B, which frequently defaulted to `open`. Qwen3-VL-4B and Qwen3-VL-8B also showed

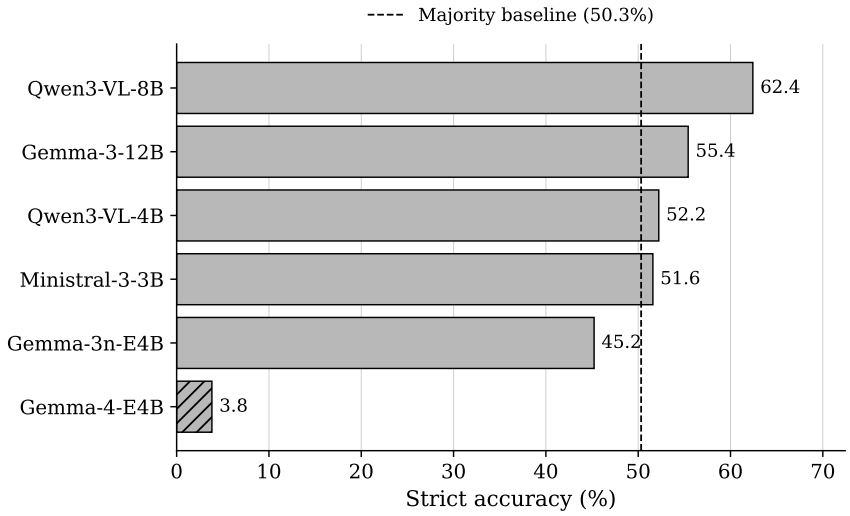

Figure 11. Strict hinge-state classification accuracy of the evaluated vision-language models on the cropped hinged-instrument test set ($N = 157$). Qwen3-VL-8B achieved the highest strict accuracy at 62.4%, followed by Gemma 3 12B at 55.4%. Most models performed only slightly above chance level, while Gemma 4 E4B performed substantially worse due to a high rate of invalid outputs under the strict evaluation criterion.

some bias toward `open`, although Qwen3-VL-8B still preserved enough discrimination to achieve the best overall result. This suggests that many models were not reliably using the subtle geometric cues needed for hinge-state recognition, but rather were defaulting to one of the states, when no conclusive answer was possible. Compared with category recognition, the effect of viewpoint was weaker and less consistent. Across the different models, no single viewpoint consistently yielded the strongest results. For the strongest model, Qwen3-VL-8B, close-up views produced the highest accuracy, followed by oblique views, while top-down views were the weakest.

Overall, the results of the hinge state classification by VLMs show that they are not accurate enough for reliable recognition in this setting. Some of the stronger models have however shown that they can capture broader visual distinctions, but they still are not comparable to vision-only models, which can run locally with much less computation power.

4.5 Additional Analysis Results

4.5.1 Grad-CAM Analysis

In order to determine which parts of the image the convolutional neural network uses to determine whether the instrument is open or closed, a technique called Gradient-weighted Class Activation Mapping (Grad-CAM) is used [SCD+19]. Grad-CAM provides interpretable explanations for predictions made by a CNN model by highlighting the regions of the image that most strongly influence a target decision. It does so by using the gradients of the target class with respect to the final convolutional layer to produce a coarse localization map. As the final layer preserves spatial information while also representing more meaningful visual patterns, Grad-CAM is useful for checking whether the classifier focuses on relevant parts of the instrument, such as the hinge or

gap between the jaws, and not irrelevant information such as the background [SCD+19].

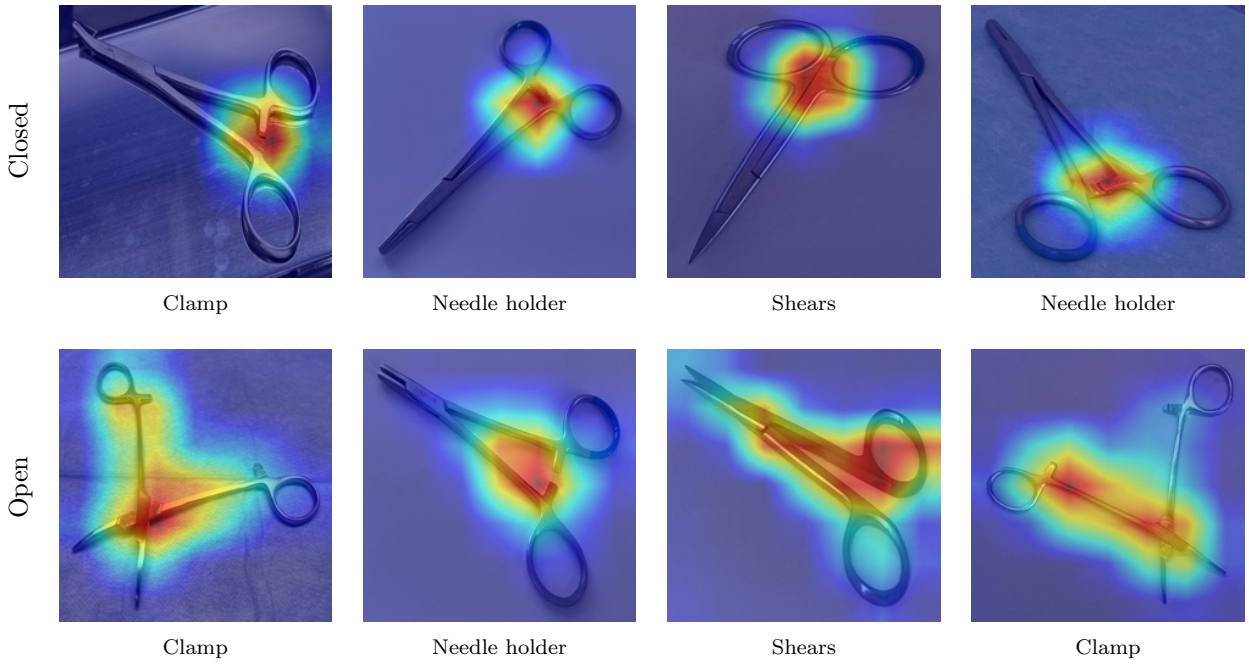

Figure 12. Grad-CAM visualizations for correctly classified hinge states. Closed examples are shown in the top row and open examples in the bottom row. In both cases, activations are concentrated around the hinge and adjacent branches, indicating that the classifier relies on relevant geometric cues for state recognition.

A small number of correctly classified images were randomly selected for Grad-CAM analysis, in which both open and closed examples were considered. As seen in Figure 12, the CNN classifiers rely on relevant geometry of the instruments to determine the open or closed states.

4.5.2 Multi-Tool Detection

Although the vision-only object detection models achieved strong performance on the standard test split, their behavior becomes more limited when multiple instruments are present in the same image. One possible reason for this behavior is the fact that the training data contained exactly one annotated instrument per image. As a result, the learned detection is optimized for single-object scenes and does not explicitly train the model to separate several nearby instruments.

Figure 13 shows that the detector is capable of identifying multiple tools in a single image to some extent, even though this setting was not part of the training setup. This suggests that learned visual features are generalized beyond the single tool per image setup used in training. However, the predictions are less reliable than in the standard test setup, and errors become more frequent.

A representative failure case is shown in Figure 14. In this example, several overlapping instruments are grouped into a single prediction rather than being separated into distinct instances. This indicates that overlap makes it difficult for the model to distinguish individual object boundaries. These scenes differ significantly from the images used for training, and can be interpreted as a limitation of the current dataset design rather than a weakness of the model architecture.

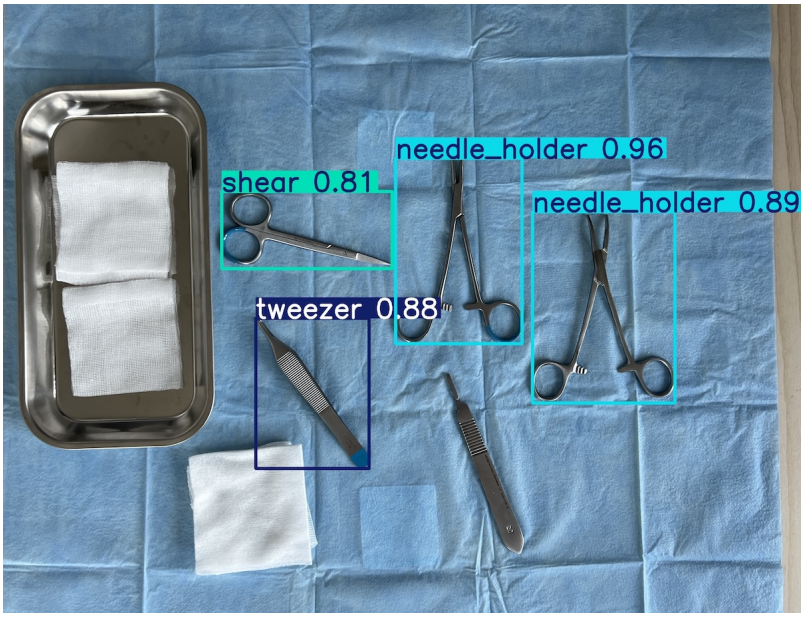

Figure 13. Example of multi-tool detection in a scene containing several instruments. The detector correctly identifies the shears, one needle holder, and the tweezers. However, the clamp is misclassified as a needle holder, and the scalpel is not detected.

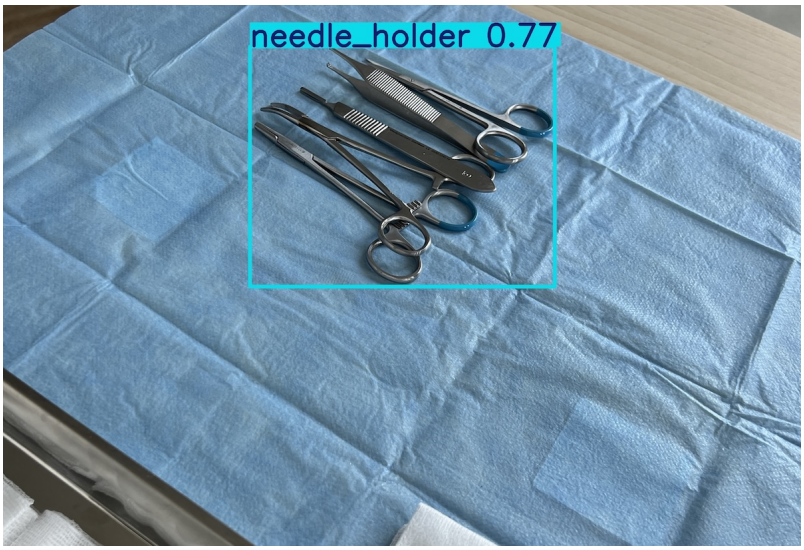

Figure 14. Failure case for overlapping multiple tools detection. Multiple instruments are grouped into a single detection and jointly classified as a needle holder, rather than being separated into distinct instances.

These examples show that the proposed system is effective for recognition of single instruments in a controlled environment, but the performance cannot be directly translated to more realistic scenes with multiple tools. Extending the dataset to include images with multiple instruments, partial occlusions, and overlapping instruments would therefore be an important future direction for research.

4.5.3 Generalization to Unseen Instances

In addition to the Grad-CAM and Multi-Tool detection experiments, the models were also evaluated on a small set of unseen instances of objects of the same class. This experiment was made as a qualitative test of generalization to new physical instances of the same broad categories. As the test set contained 33 images across 11 scenes and 3 viewpoints, results should be interpreted conservatively.

As seen in Table 14, detection of instruments in the image generalized well. Both YOLO26 models detected an object in all 33 images. At the coarse level of distinguishing hinged instruments from non-hinged instruments, both detection models were fully correct on this test. The main generalization issue was therefore not object localization on the image, but fine-grained category classification.

Table 14. Detector performance on the unseen-instance set.

Model	Detection rate	Tool accuracy	Hinged/non-hinged accuracy	Mean confidence
YOLO26n	33/33 (100.0%)	20/33 (60.6%)	33/33 (100.0%)	0.82
YOLO26s	33/33 (100.0%)	24/33 (72.7%)	33/33 (100.0%)	0.92

More specifically, the main failure mode was confusion between needle holders and clamps. For both detectors, shears and tweezers were recognized correctly in all images, whereas needle holders were less accurate. YOLO26s performed better overall than YOLO26n, but both models misclassified needle holders as clamps. This suggests that the general representation of the instruments transferred reasonably well to most unseen categories.

One important observation that can be made is that the misclassifications of instruments appear to be a systematic error rather than random error. For YOLO26s, several unseen needle holder scenes were misclassified as clamps in all three viewpoints of the same scene. These mistakes were also made with high confidence, which indicates that high confidence does not reliably indicate successful generalization.

In order to investigate the hinge-state recognition, the cropped detections were also classified using the hinge-state classifiers. Table 15 reports the best performing classifier (EfficientNet-B0) combined with each detection model.

The results demonstrate that hinge-state recognition is better generalized than fine-grained tool category recognition in this dataset. The classification model was able to correctly predict the open/closed state in 26 out of 27 cases. However, end-to-end state accuracy, which includes both state classification and category recognition, dropped substantially. This indicates that the main

Table 15. Best hinge-state classification results on the unseen-instance set (EfficientNet-B0).

Pipeline	State accuracy on hinged crops	End-to-end state accuracy
YOLO26n + EfficientNet-B0	26/27 (96.3%)	14/27 (51.9%)
YOLO26s + EfficientNet-B0	26/27 (96.3%)	18/27 (66.7%)

limitation of generalization in this experiment was due to the category confusion of the detector, and not the binary state classifier.

Overall, this experiment suggests that the proposed pipeline has the ability to a certain extent to generalize on unseen instrument instances. As the sample size of the experiment was limited and did not include examples of all five classes, these findings should be treated as indicative. They nevertheless demonstrate that fine-grained category recognition is more sensitive to instance variation than either detection or hinge-state classification.

5 Discussion

This thesis investigated how an active imaging setup can be designed to reliably recognize handheld surgical instruments and their hinge state under controlled variation in viewpoint, background, and lighting. The results show that such a system can be effectively designed in a setting with a fixed set of predefined classes. The combination of a vision-only detector with a hinge-state classifier particularly proved to be effective for the five instrument categories considered in this study. Both YOLO26 detection models achieved very strong performance on the main test split, while all four hinge-state classifiers performed strongly on the cropped hinged instrument images, with EfficientNet-B0 achieving the highest overall accuracy.

It is important to consider that the main contribution of this study is a system design and evaluation contribution rather than a new model. The thesis does not propose model architecture novelty or propose a fundamentally new algorithm. Instead, it demonstrates how a complete active vision pipeline can be designed, implemented, and evaluated end-to-end for a specific problem. This includes dataset design, data acquisition, annotation, model selection, pipeline integration, evaluation, simulation of active multi-view acquisition, and qualitative analysis.

The results of this study first show that vision-only models are highly effective under the closed-world assumption considered in this study. For object detection, both YOLO26n and YOLO26s reached very high precision, recall, and mAP scores. This indicates that the five instruments can be localized on the image, and classified correctly under the conditions of the dataset. This is consistent with previous work showing that CNN-based and more specifically YOLO-based approaches are able to achieve strong performance for surgical instrument recognition, especially in cases where the recognition environment is constrained and the target categories are clearly defined [LKB+23, XLC+25]. This result should however be interpreted taking into account that the dataset was acquired in a highly controlled environment. Prior work on open surgery emphasizes that surgical tool detection is particularly more challenging as there are variable imaging conditions, complex occlusions with hands and other instruments, as well as a lack of large annotated datasets [FHKS22, SHK+21]. The high performance that is observed in this study therefore demonstrates the feasibility

of the proposed pipeline under specific conditions, and does not imply that the same performance would transfer directly to realistic surgical scenes.

One central finding of the study is that viewpoint influenced the two recognition tasks differently. For object detection, close-up views resulted in the strongest overall performance for both YOLO models. This suggests that tool category recognition benefits most from views in which local details such as jaw shape, blade structure, tip geometry, and other fine differences between similar instruments are visible. This is especially relevant to visually similar categories such as clamps and needle holders, where the distinction between the two is made not on the global shape of the instrument but rather the fine details such as the shape of the tip, or jaw pattern. This interpretation is aligned with prior work on multi-camera surgical tool recognition, where close-up views were described as useful for distinguishing similar tools [BGPL22]. It is also supported by recent work on ultra-fine-grained surgical instrument classification, which emphasizes that visually similar instruments may differ only slightly in design and that detailed views of the instrument tips, curvature, surface texture, and tooth patterns are important for classification [ADMA+25].

In contrast, hinge-state classification showed a different pattern. For EfficientNet-B0, ResNet-18, and YOLO26n-cls, the top down view yielded the highest accuracy, while close-up views were generally the weakest. This suggests that unlike instrument detection, hinge-state classification depends less on the detailed local information, but more on the global geometry of the instrument, such as the separation between the two branches and the angle at the hinge. This supports the distinction made by Gouidis et al. between object detection and object-state detection, where object states are treated as a separate visual recognition problem from object categories. Their work also shows that state recognition can rely on different cues than those used for object categorization [GPAP21]. Grad-CAM analysis further supports this as it has shown to consistently focus on the hinge part of the instrument when determining the open or closed state. This difference between category recognition and state recognition shows that viewpoint selection should not be done in a task-agnostic way. In other words, a viewpoint that is useful for identifying the instrument category is not necessarily the best viewpoint for determining the state of the instrument. This supports the motivation for active and multi-view systems. Prior work has also shown that multi-camera setups can improve robustness in tool recognition in open surgery contexts by reducing failures caused by occlusions, or by the loss of visibility by a single camera perspective [BGPL22]. The results of this thesis add to this discussion by showing that different viewpoints may be useful for different recognition goals.

The same-scene multi-view experiment further shows the value of multi-view and active perception, but also highlights an important accuracy-efficiency trade-off. In this experiment on the extended dataset, adding viewpoints generally improved performance over relying only on the initial top-down view. However, the active policy did not consistently outperform full fusion of all three views. Instead, the best active configuration matched the strongest full-fusion end-to-end result, while other active configurations used fewer views at the cost of slightly lower accuracy. This suggests that the added value of active acquisition is not simply that it always maximizes accuracy, but that it allows the system to adapt the number of acquired views based on prediction confidence. Previous work on efficient multi-view understanding has also shown that selecting informative views can be preferable to processing all available views indiscriminately [HGZ24]. In this thesis, the threshold-based active image acquisition strategy therefore provided a simple rule-based implementation of active view

acquisition.

The comparison between YOLO26n and YOLO26s also shows that model selection should be done with consideration to the system as a whole rather than solely on performance. The two models performed very similarly overall, but YOLO26n achieved a slightly higher mAP@50:95 and also had lower inference latency, whereas YOLO26s had slightly higher precision and recall. This suggests that smaller models may be preferred when responsiveness, efficiency and low latency are requirements of the problem. This is especially true in this application as inference is performed repeatedly as new views are acquired. The larger model may still be a better choice when some detection metrics, or better generalization performance is prioritized. Therefore, the choice of detector should depend not only on the accuracy but also on constraints of the system.

The hinge-state classification results also provide interesting perspectives on the difficulty of state recognition. It may seem that the high accuracies achieved on all four classifiers suggest that the task is simple. However there are clear differences in the recall performance of closed and open classes, where closed classes have consistently higher recall rates. A plausible explanation is that the closed class is visually more consistent, whereas the open class contains a wider range of valid opening angles, and therefore more variability within the class.

The comparison of VLMs with vision-only models provides a clear answer on whether VLMs are viable alternatives to vision-only models. In the setup of this study, VLMs did not show any meaningful improvement over vision-only models. Although the best VLMs performed moderately well on the closed-set instrument category recognition task, their performance remained well below that of the trained vision-only detectors. Their performance on the hinge-state classification task was extremely limited: the best performing models gave results slightly above chance level for a binary task, and inference latency was also much higher than for the vision-only models. This latency difference is particularly relevant for an active multi-view system where predictions must be repeated after each newly acquired view.

Prompt optimization improved performance for some models, but it was not sufficient to perform comparatively to vision-only models. These findings suggest that, at least in the present zero-shot setting, general-purpose VLMs are not yet viable for fine-grained surgical instrument recognition. This is especially true for hinge-state classification, where the task depends on local geometric differences. This is consistent with recent discussions of VLMs in medical and specialized visual domains, where limited domain-specific data, and fine-grained visual recognition are still important challenges [LWD+25, SFB+26]. Although specialized VLM approaches have shown promising results for promptable surgical vision tasks such as instrument segmentation [ZAW+23], the results of this thesis suggest that zero-shot prompting of general purpose VLMs is not sufficient for reliable surgical instrument recognition.

The qualitative analysis of the results provides some supporting evidence. The Grad-CAM visualizations indicate that the hinge-state classifier does rely on meaningful local geometry, as activations are concentrated around the hinge region and the branches rather than on background noise. This strengthens the claim that the classifier is learning geometric cues that are relevant to the classification task rather than using obvious background artifacts.

The largest limitations of this study include the dataset realism. The setup was a controlled environment, and the experiments were conducted under a closed-world assumption where the

categories are predefined and the training and test data belong to the same setup. This was an appropriate setup for isolating the effect of viewpoint and for building a first active recognition pipeline, but it also limits the extent to which this system can be generalized to real surgical environments. In realistic open surgery settings, instruments may be partially occluded, held by hands, surrounded by tissue or other objects, and observed under more variable lighting and camera conditions [FHKS22, SHK+21, XLC+25]. The test images also contain only one instrument per image, and the backgrounds, and environment were designed for this specific experiment, and did not naturally occur in a setting where surgical instruments are usually found. Furthermore, the size of the full dataset is relatively small, and the viewpoint subsets for hinge-state classification were not perfectly balanced.

The multi-tool experiment illustrates the limitations of the dataset clearly. Although the detector was able to identify multiple tools in some cases, its performance dropped when instruments overlapped or appeared close together. In the most difficult failure cases, several overlapping tools were grouped into a single detection. This should not be interpreted as a weakness of the detector model, but rather as a weakness in the dataset and training design. The detector was trained only on single objects per scene, and therefore was not optimized for separating multiple nearby instances of surgical instruments. The multi-tool experiment therefore reveals a limitation of the dataset design and task formulation. Extending the dataset to more realistic scenes with multiple instruments, partial occlusions, overlap, and more natural obstacles would therefore be an important next step to reinforce this current setup.

The generalization experiment further shows that model robustness to unseen instances of instruments is a challenge. Both YOLO detectors localized instruments reliably in the generalization set, and were also correctly distinguishing between hinged and non-hinged instruments. However, fine-grained category recognition yielded weaker results, especially for unseen needle holders that were misclassified as clamps. This suggests that the models learned features that can be transferred to some extent for broad categories, but that fine-grained category recognition is sensitive to specific instances. The unseen generalization experiment was limited in dataset size, and number of tested instances. The findings should therefore be interpreted as indicative.

There are several directions for future work that can be concluded from these limitations. First, the dataset can be extended to more realistic scenes and conditions. This can include multiple instruments per image, more varied occlusions, or a greater environmental variation. Second, the current active acquisition strategy is based on rules and thresholds. Future work can investigate how this active selection strategy can be done more intelligently using for example reinforcement learning to decide which next view is more informative. Third, the generalization of the system to unseen instances should be studied more systematically, with a greater number of instances, and on different datasets. Finally, VLMs could be further studied by investigating newer types of specialized VLMs that have been trained on medical data, or using custom fine-tuned VLMs, rather than zero-shot prompting that was used in this study.

Overall, the results of this thesis show that active instrument recognition is highly feasible in a controlled environment, and that taking into account viewpoint is important in the design of a system to achieve strong performance. The study demonstrates that different recognition tasks benefit from different viewpoints, that selective acquisition of views can provide a trade-off between accuracy and efficiency, and that specialized vision-only models remain the strongest option for

this problem currently. The findings also show that robustness of the system in more realistic environments will not only depend on model choice, but also on richer training data, including more realistic scenes and more advanced perception strategies.

6 Conclusion

The thesis designed and evaluated an active vision system for recognizing handheld surgical instruments and their hinge state. Using a balanced dataset with controlled variation in viewpoint and background, the proposed system combined vision-only detection models with hinge-state classification and evaluated this pipeline using single view, multi-view, active acquisition, and zero-shot vision language models.

With respect to the Main RQ, the results show that such a system can indeed be designed in an effective manner in a setting with a fixed predefined set of classes. In particular, the combination of a vision-only object detector with a hinge-state classifier proved highly effective for the five instrument categories considered in this study. Both YOLO26 detection models tested achieved very strong performance on the main test split, while all four hinge-state classifiers performed strongly on cropped hinged instrument images, with EfficientNet-B0 achieving the highest overall accuracy.

With respect to RQ1, the results show that the most informative viewpoint depends on the recognition task. For instrument category recognition, close-up views were in general the most discriminative, while for the hinge-state classification, top-down views produced the strongest results for most models. This suggests that category recognition benefits from detailed local visual information such as tip geometry, while hinge-state classification depends more on the global geometric configuration of the instrument, such as the angle at the hinge and the separation of the branches.

With respect to RQ2, the proposed vision-only system reached high recognition performance under the controlled conditions of the experiment in this study. Both YOLO26 detection models performed strongly across the five instrument categories, and all evaluated hinge-state classifiers achieved high accuracy on cropped images of hinged instruments, with EfficientNet-B0 performing best overall. These findings show that reliable real-time recognition of both instrument category and open or closed hinge state is feasible in a controlled closed-world setting.

With respect to RQ3, adaptive multi-view acquisition improved robustness compared with relying on a single initial top-down view. In the same-scene experiments, the strongest active configuration achieved 0.733 end-to-end hinge-state accuracy using 2.50 views on average, compared with 0.550 for top-only inference and 0.700 for full fusion using the same detector-classifier pair. These results show that adaptive acquisition can provide strong performance without always requiring fusion of all available views.

With respect to RQ4, the comparison with vision-language models further showed that zero-shot multimodal models are not competitive with specialized vision-only models for this task. Although some VLMs achieve moderate results on tool category recognition, their performance remained below that of the trained vision-only pipeline, especially for hinge state classification, and also had a much higher inference latency. These findings indicate that task-specific models remain the more

suitable choice for fine-grained instrument recognition.

Some limitations of this study include that the experiments were conducted under strict acquisition conditions and within a closed-world assumption with predefined classes. The dataset contained only one instrument per image, which limited the ability of the system to generalize to more realistic scenes that contain multiple overlapping tools, clutter, or larger occlusions.

Future work should therefore focus on extending both the training dataset and the setup to more realistic scenarios. This includes collecting images with multiple instruments per scene, stronger occlusion, and more natural surgical backgrounds. In addition, the current threshold-based active acquisition strategy could be replaced by a reinforcement learning policy that determines more intelligently which next viewpoint is most informative. Finally, further work could investigate whether multimodal models such as VLMs become more useful when combined with fine-tuning, or domain-specific adaptation. Overall, this thesis shows that active surgical instrument recognition is promising, and that careful system design is central to achieving reliable performance.

References

[ADMA+25] Md. Atabuzzaman, Gino Di Matteo, Hani Alomari, Chiawei Tang, Connor Hale, Adam E. Goode, David Ryan King, and Chris Thomas. Real-time ultra-fine-grained surgical instrument classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2070–2079, June 2025.

[ASLLJ17] Jean-Baptiste Alayrac, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *International Conference on Computer Vision (ICCV)*, 2017.

[BGPL22] Kristina Basiev, Adam Goldbraikh, Carla M. Pugh, and Shlomi Laufer. Open surgery tool classification and hand utilization using a multi-camera system. *International Journal of Computer Assisted Radiology and Surgery*, 17(8):1497–1505, August 2022.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[FHKS22] Ryo Fujii, Ryo Hachiuma, Hiroki Kajita, and Hideo Saito. Surgical tool detection in open surgery videos. *Applied Sciences*, 12(20):10473, 2022.

[Goo25a] Google. Gemma 3 12B. https://huggingface.co/google/gemma-3-12b-it, 2025. Hugging Face model card. Accessed: 2026-04-10.

[Goo25b] Google. Gemma 3n E4B. https://huggingface.co/google/gemma-3n-E4B, 2025. Hugging Face model card. Accessed: 2026-04-10.

[Goo26] Google. Gemma 4 E4B. https://huggingface.co/google/gemma-4-E4B, 2026. Hugging Face model card. Accessed: 2026-04-10.

[GPAP21] Filippos Gouidis, Theodoris Patkos, Antonis A. Argyros, and Dimitris Plexousakis. Detecting object states vs detecting objects: A new dataset and a quantitative experimental study, 2021. arXiv:2112.08281.

[HGZ24] Yunzhong Hou, Stephen Gould, and Liang Zheng. Learning to select views for efficient multi-view understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20135–20144, June 2024.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[KHDM98] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[KHE+25] Linus L. Kienle, Anna Hilsmann, Peter Eisert, Michael Knoke, and Eric L. Wisotzky. Surgical instrument detection on the instrument stand using neural networks. *Current Directions in Biomedical Engineering*, 11(1):532–535, 2025.

[LCHW21] Jiann-Der Lee, Jong-Chih Chien, Yu-Tsung Hsu, and Chieh-Tsai Wu. Automatic surgical instrument recognition—a case of comparison study between the faster R-CNN, mask R-CNN, and single-shot multi-box detectors. *Applied Sciences*, 11(17):8097, 2021.

[LKB+23] Jan Lehr, Kathrin Kelterborn, Clemens Briese, Marian Schlueter, Ole Kroeger, and Joerg Krueger. Image-based recognition of surgical instruments by means of convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 18(11):2043–2049, November 2023.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[LS22] Margret Liehn and Hannelore Schlautmann. *101 of Surgical Instruments: Naming, Recognizing, Handling and Presenting*. Springer, Berlin, Germany, 2022.

[LWD+25] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025. arXiv:2501.02189.

[LYP+24] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models, 2024. arXiv:2312.07533.

[MAS+22] Pietro Mascagni, Deepak Alapatt, Luca Sestini, Maria S. Altieri, Amin Madani, Yusuke Watanabe, Adnan Alseidi, Jay A. Redan, Sergio Alfieri, Guido Costamagna, Ivo Boskoski, Nicolas Padoy, and Daniel A. Hashimoto. Computer vision in surgery: From potential to clinical value. *npj Digital Medicine*, 5(1):163, 2022.

[Mis25] Mistral AI. Ministral 3 3B Instruct 2512. https://huggingface.co/mistralai/Ministral-3-3B-Instruct-2512, 2025. Hugging Face model card. Accessed: 2026-04-10.

[Qwe25a] Qwen Team. Qwen3-VL-4B-Instruct. https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct, 2025. Hugging Face model card. Accessed: 2026-04-10.

[Qwe25b] Qwen Team. Qwen3-VL-8B-Instruct. https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct, 2025. Hugging Face model card. Accessed: 2026-04-10.

[RDGF16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[RRP24] Tobias Rueckert, Daniel Rueckert, and Christoph Palm. Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art. *Computers in Biology and Medicine*, 169:107929, 2024.

[SCD+19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[SFB+26] Kirill Skobelev, Eric Fithian, Yegor Baranovski, Jack Cook, Sandeep Angara, Shauna Otto, Zhuang-Fang Yi, John Zhu, Daniel A. Donoho, X. Y. Han, Neeraj Mainkar, and Margaux Masson-Forsythe. A comparative study in surgical AI: Datasets, foundation models, and barriers to Med-AGI. *medRxiv*, 2026.

[SHK+21] Takafumi Shimizu, Ryo Hachiuma, Hiroki Kajita, Yuki Takatsume, and Hideo Saito. Hand motion-aware surgical tool localization and classification from an egocentric camera. *Journal of Imaging*, 7(2):15, 2021.

[Sze22] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer, Cham, 2 edition, 2022.

[TL19] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

[TMHL25] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software. https://github.com/HumanSignal/label-studio, 2020–2025. Open-source software. Accessed: 2026-04-10.

[Tor26] TorchVision. Models and pre-trained weights. https://docs.pytorch.org/vision/stable/models.html, 2026. Documentation page. Accessed: 2026-04-10.

[Ult25a] Ultralytics. Image classification with Ultralytics YOLO. https://docs.ultralytics.com/tasks/classify/, 2025. Documentation page. Accessed: 2026-04-10.

[Ult25b] Ultralytics. Object detection with Ultralytics YOLO. https://docs.ultralytics.com/tasks/detect/, 2025. Documentation page. Accessed: 2026-04-10.

[Ult25c] Ultralytics. Ultralytics YOLO26. https://docs.ultralytics.com/models/yolo26/, 2025. Documentation page. Accessed: 2026-04-10.

[XLC+25] Zhaokun Xu, Feng Luo, Feng Chen, Hang Wu, and Ming Yu. Surgical tool detection in open surgery based on improved-YOLOv8. *Biomedical Signal Processing and Control*, 105:107548, 2025.

[ZAW+23] Zijian Zhou, Oluwatosin Alabi, Meng Wei, Tom Vercauteren, and Miaojing Shi. Text promptable surgical instrument segmentation with vision-language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 28611–28623. Curran Associates, Inc., 2023.

[ZW17] Tian Zhou and Juan P. Wachs. Needle in a haystack: Interactive surgical instrument recognition through perception and manipulation. *Robotics and Autonomous Systems*, 97:182–192, 2017.

Appendix A Surgical Instruments

Below is the list of instruments that have been used in this research. Main instruments were used for the creation of the training, validation, and testing sets of all vision models.

Table A.0.1. List of surgical instruments used in the main and generalization experiments.

Instrument description	Category	Experiment	Image
Mayo-Hegar needle holder (15.5 cm)	NH	Main	
Adson tissue forceps (12.5 cm)	TW	Main	
Iris scissors, straight (11 cm)	SH	Main	
Mosquito hemostatic forceps	CM	Main	
Scalpel handle No.3	SC	Main	
Disposable needle holder (12 cm)	NH	Generalization	
Disposable scissors (11.6 cm)	SH	Generalization	
Disposable anatomical forceps (12 cm)	TW	Generalization	

Appendix B Dataset

B.1 Full Dataset

The dataset is publicly available on Hugging Face for reproducibility and further research.[1]

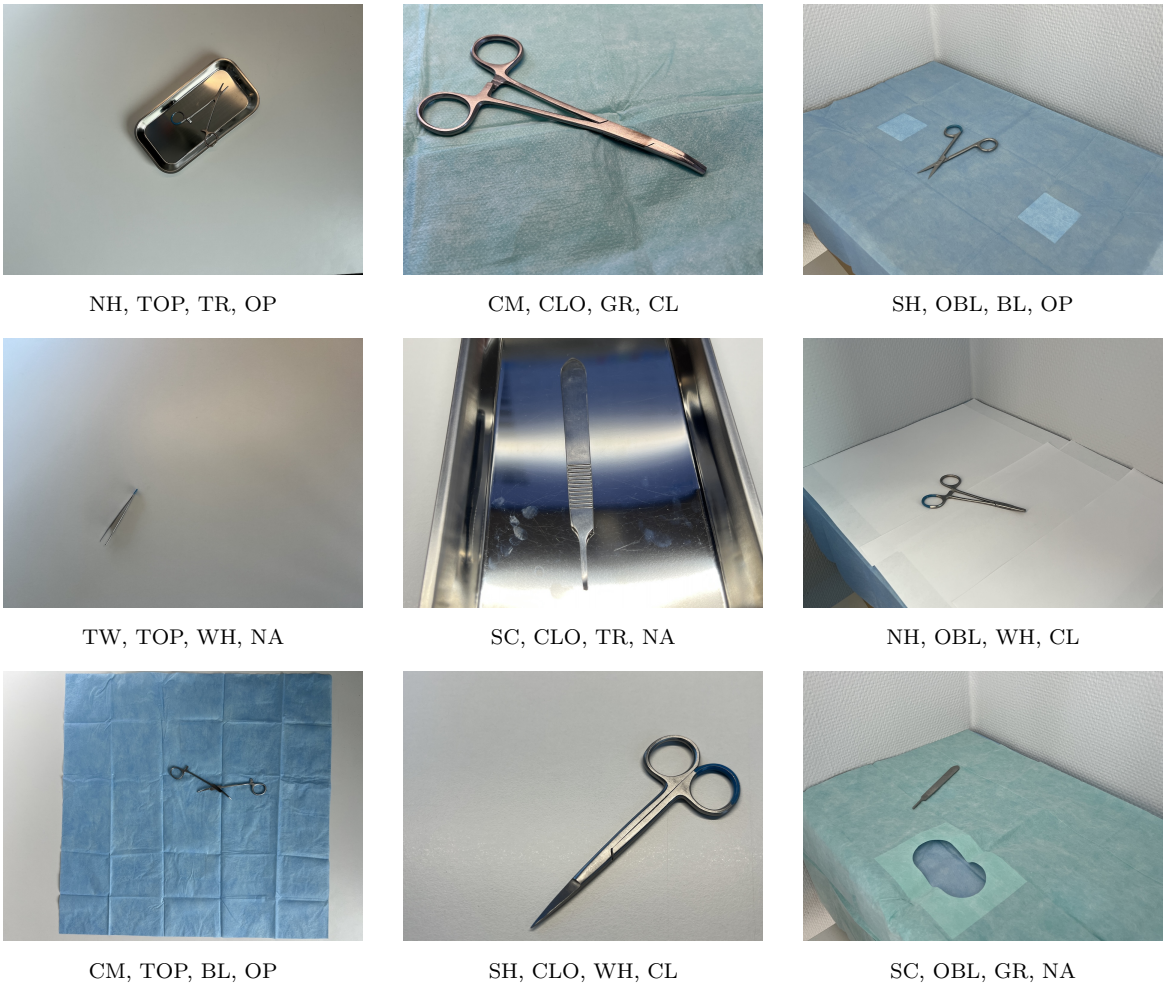

Figure B.1.1. Representative samples from the full surgical instrument dataset. Labels indicate instrument category, viewpoint, background, and state.

[1] https://huggingface.co/datasets/joonhaim/surgical-tool-recognition-full-multiview

Table B.1.1. Summary statistics of the collected surgical instrument image full dataset.

Property	**Value**
Total images	2520
No. of classes	5
Classes	NH, SH, TW, SC, CM
No. of viewpoints	3
Viewpoints	TOP, OBL, CLO
No. of backgrounds	4
Backgrounds	TR, WH, BL, GR
Images per class	504
Images per viewpoint	840
Images per background	630
Class–view–background groups	60
Images per group	42
Hinged classes	3 (NH, SH, CM)
Non-hinged classes	2 (TW, SC)
Hinged images	1512
Non-hinged images	1008
Open images (OP)	756
Closed images (CL)	756
NA state images	1008
Images per hinged-state folder	21
Images per non-hinged folder	42
Dataset size	7.39 GB

Table B.1.2. Per-class distribution of the full dataset split used for object detection.

Class	**Train**	**Validation**	**Test**	**Total**
CM	397	51	56	504
NH	404	47	53	504
SC	405	58	41	504
SH	408	48	48	504
TW	402	48	54	504
Total	2016	252	252	2520

B.2 Hinge Classifier Cropped Dataset

The dataset is publicly available on Hugging Face for reproducibility and further research.[2]

Figure B.2.1. Representative samples from the cropped hinge-state classification dataset.

[2] https://huggingface.co/datasets/joonhaim/hinged-state-classifier-crops

Table B.2.1. Summary statistics of the cropped dataset used for hinged instrument state classification.

Property	**Value**
Total cropped images	1512
Task	Binary state classification
States	open, closed
Source instrument classes	3 (CM, NH, SH)
Images per source class	504
Total open images	756
Total closed images	756
Train images	1209
Validation images	146
Test images	157
Global tool distribution	CM: 504, NH: 504, SH: 504
Global viewpoint distribution	CLO: 504, OBL: 504, TOP: 504
Global background distribution	BL: 378, GR: 378, TR: 378, WH: 378
Dataset size	1.14 GB

Table B.2.2. Per-split distribution of the cropped hinged-instrument state classification dataset.

Split	**Total**	**Open**	**Closed**	**CM**	**NH**	**SH**	**CLO**	**OBL**	**TOP**	**BL**	**GR**	**TR**	**WH**
Train	1209	603	606	397	404	408	405	404	400	300	307	299	303
Val	146	75	71	51	47	48	52	54	40	40	25	42	39
Test	157	78	79	56	53	48	47	46	64	38	46	37	36
Total	1512	756	756	504	504	504	504	504	504	378	378	378	378

B.3 Extended Dataset: Linked Scenes

TOP OBL CLO

Figure B.3.1. Example linked scene from the extended dataset (Scene S009, needle holder, open), showing the same instrument configuration captured from the top-down, oblique, and close-up viewpoints. Many of these scenes are challenging, including partial occlusions.

Table B.3.1. Summary of the extended linked-scenes dataset used for active multi-view evaluation.

Property	**Value**
Total images	300
Total scenes	100
Scene structure	Each scene captured from 3 viewpoints
Classes	CM, NH, SC, SH, TW
Images per class	60 images per class
Scenes per class	20 scenes per class
Viewpoints	TOP, OBL, CLO
Images per viewpoint	100 images per viewpoint
States	CL, OP, NA
Images per state	CL: 90; OP: 90; NA: 120
Scenes per state	CL: 30; OP: 30; NA: 40
Dataset size	960.9 MB

B.4 Generalization Test Dataset

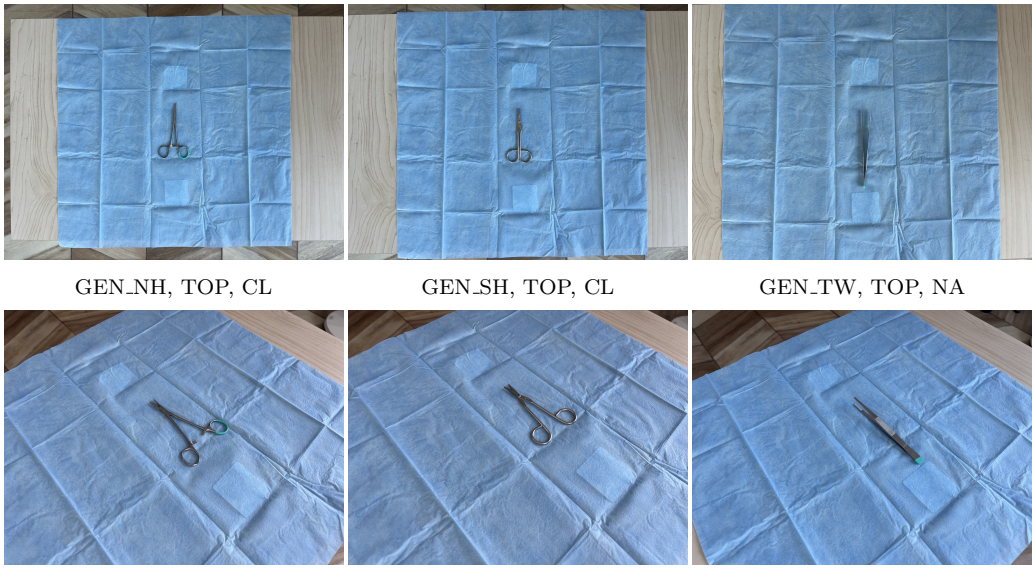

Figure B.4.1. Representative samples from the generalization test dataset. These images show previously unseen physical instrument instances from the same predefined classes as in the main dataset.

Table B.4.1. Summary of the generalization test dataset.

Property	**Value**
Total images	33
Scene structure	11 scenes, each captured from 3 viewpoints
Classes	NH, SH, TW
Scenes per class	NH: 6; SH: 3; TW: 2
Images per class	NH: 18; SH: 9; TW: 6
Viewpoints	TOP, OBL, CLO
States	CL, OP, NA
Images per state	CL: 9; OP: 18; NA: 6
Dataset size	115.46 MB

Appendix C VLM Prompts

C.1 Instrument Recognition Base Prompt

System Prompt (Base Case)

```
You are given one image containing a single main surgical instrument.
Classify the instrument into exactly one of the allowed categories.

Allowed categories:
clamp
needle_holder
scalpel
shear
tweezer

Rules:
- Return exactly one label from the allowed categories.
- Do not explain your answer.
- Do not output JSON.
- Do not output punctuation.
- Do not output any extra words.
- If uncertain, return the single most likely category.
```

User Prompt (Base Case)

```
Identify the surgical instrument shown in the image.

Return exactly one of:
clamp
needle_holder
scalpel
shear
tweezer
```

C.2 Instrument Recognition Optimized Prompt

System Prompt (Optimized)

```
You are performing closed-set classification on one image containing a single main
    surgical instrument.

Choose exactly one label from the following allowed categories:
clamp
needle_holder
scalpel
shear
tweezer

Visual descriptions of the categories:
- clamp: ring-handled hinged instrument with long shanks and blunt jaws; jaws are curved;
    the tip looks like a gripping end rather than a cutting blade
- needle_holder: ring-handled hinged instrument, straight shanks, and short, sturdy, blunt
     jaws. The working end looks compact, thick, and rectangular rather than long, pointed,
     or blade-like.
- scalpel: flat straight metal handle without finger rings; elongated body with a narrow
    end for mounting a blade; may appear with no blade attached
- shear: ring-handled instrument with two long opposing cutting blades; blades are narrow
    and pointed or fine-tipped; overall appearance resembles surgical scissors
- tweezer: instrument without finger rings, formed by two spring-like arms joined at one
    end; narrow pinching tips and a flat textured gripping area in the middle

Decision guidance:
- Focus only on the instrument's visible shape.
- Use handle type, hinge structure, overall silhouette, and tip or blade shape.
- Ring handles suggest clamp, needle_holder, or shear.
- No ring handles and a flat straight handle suggest scalpel.
- No ring handles and two spring-like arms suggest tweezer.
- If the instrument has two visible cutting blades, classify it as shear.
- If the instrument has blunt jaws instead of blades, classify it as clamp or
    needle_holder.
- If uncertain between clamp and needle_holder, prefer needle_holder only when the jaws
    look shorter, thicker, and more compact.


Output rules:
- Return exactly one label from the allowed categories.
- Output only the label.
- Do not explain your answer.
- Do not output JSON.
- Do not output punctuation.
- Do not output any extra words.
- Do not use synonyms.
- If uncertain, return the single most likely label.
```

User Prompt (Optimized)

```
Identify the surgical instrument shown in the image.

Return exactly one of:
clamp
needle_holder
scalpel
shear
tweezer
```

C.3 Hinge Classification Prompt

System Prompt

```
You are given one crop of a single hinged surgical instrument.
Classify the hinge state into exactly one of the allowed categories.

Allowed categories:
open
closed

Rules:
- Return exactly one label from the allowed categories.
- Do not explain your answer.
- Do not output JSON.
- Do not output punctuation.
- Do not output any extra words.
- If uncertain, return the single most likely category.
```

User Prompt

```
Identify the hinge state of the surgical instrument crop.

Return exactly one of:
open
closed
```

Appendix D Detailed Results

D.1 Vision-Only Detection Models

Table D.1.1. Overall object detection performance on the test split ($N = 252$ images) for the two YOLO26 detection models.

Model	Precision	Recall	mAP@50	mAP@50–95
YOLO26n	0.9854	0.9842	0.9945	0.9816
YOLO26s	0.9892	0.9853	0.9936	0.9807

Table D.1.2. Class-wise object detection results on the test split for YOLO26n and YOLO26s. Since each image contains one annotated instrument, N denotes both the number of images and instances.

Model	Class	N	Precision	Recall	F1	mAP@50	mAP@50–95
YOLO26n	Clamp	56	1.0000	0.9626	0.9809	0.9943	0.9862
YOLO26n	Needle holder	53	0.9868	0.9811	0.9840	0.9948	0.9907
YOLO26n	Scalpel	41	0.9647	1.0000	0.9820	0.9945	0.9719
YOLO26n	Shears	48	0.9757	1.0000	0.9877	0.9942	0.9933
YOLO26n	Tweezers	54	1.0000	0.9773	0.9885	0.9948	0.9661
YOLO26s	Clamp	56	0.9821	0.9773	0.9796	0.9938	0.9854
YOLO26s	Needle holder	53	1.0000	0.9806	0.9902	0.9948	0.9894
YOLO26s	Scalpel	41	0.9697	1.0000	0.9846	0.9950	0.9741
YOLO26s	Shears	48	0.9942	1.0000	0.9971	0.9950	0.9898
YOLO26s	Tweezers	54	1.0000	0.9686	0.9841	0.9893	0.9648

D.2 Vision-Only Hinge Classification Models

Table D.2.1. Detailed hinge-state classification results on the test set ($N = 157$). OP denotes the open state and CL denotes the closed state. P, R, and F1 denote precision, recall, and F1-score, respectively.

Model	Acc.	Open			Closed			Confusion counts			
		P	R	F1	P	R	F1	CL→CL	CL→OP	OP→CL	OP→OP
EfficientNet-B0	98.09	100.00	96.15	98.04	96.34	100.00	98.14	79	0	3	75
YOLO26n-cls	97.45	98.68	96.15	97.40	96.30	98.73	97.50	78	1	3	75
ResNet-18	97.45	100.00	94.87	97.37	95.18	100.00	97.53	79	0	4	74
YOLO26s-cls	96.82	98.67	94.87	96.73	95.12	98.73	96.89	78	1	4	74

D.3 Active-Vision Results

Table D.3.1. Scene-level TOP-only and full-fusion multi-view results on the extended same-scene dataset. Tool accuracy is reported over all 100 scenes. Conditional and end-to-end state accuracy are reported for the 60 hinged scenes.

		TOP-only			Full fusion		
Detector	**Classifier**	**Tool**	**Cond. state**	**E2E state**	**Tool**	**Cond. state**	**E2E state**
YOLO26n	EfficientNet-B0	0.550	0.917	0.550	0.830	0.824	0.700
YOLO26n	ResNet-18	0.550	0.833	0.500	0.830	0.824	0.700
YOLO26n	YOLO26n-cls	0.550	0.750	0.450	0.830	0.804	0.683
YOLO26n	YOLO26s-cls	0.550	0.806	0.483	0.830	0.784	0.667
YOLO26s	EfficientNet-B0	0.790	0.889	0.667	0.850	0.880	0.733
YOLO26s	ResNet-18	0.790	0.822	0.617	0.850	0.860	0.717
YOLO26s	YOLO26n-cls	0.790	0.800	0.600	0.850	0.860	0.717
YOLO26s	YOLO26s-cls	0.790	0.844	0.633	0.850	0.840	0.700

Table D.3.2. Best active multi-view policy results after threshold sweeping on the extended same-scene dataset. Mean views denotes the average number of acquired views per scene.

		Best thresholds		Best active policy			
Detector	**Classifier**	τ_{tool}	τ_{state}	**Tool**	**Cond. state**	**E2E state**	**Mean views**
YOLO26n	EfficientNet-B0	0.90	0.50	0.830	0.863	**0.733**	2.500
YOLO26n	ResNet-18	0.90	0.50	0.830	0.843	0.717	2.500
YOLO26n	YOLO26n-cls	0.90	0.70	0.830	0.804	0.683	2.520
YOLO26n	YOLO26s-cls	0.90	0.50	0.830	0.843	0.717	2.500
YOLO26s	EfficientNet-B0	0.40	0.65	0.820	0.894	0.700	1.320
YOLO26s	ResNet-18	0.40	0.95	0.820	0.851	0.667	1.480
YOLO26s	YOLO26n-cls	0.95	0.75	0.820	0.894	0.700	2.410
YOLO26s	YOLO26s-cls	0.70	0.50	0.820	0.851	0.667	1.670

D.4 VLM Instrument Recognition

Table D.4.1. Overall and class-wise VLM tool-category classification results on the test set ($N = 252$).

(a) Overall classification performance

Model	Correct	Strict acc. (%)	Invalid	Valid acc. (%)	Avg. lat. (s)
Base prompt					
Gemma-3-12B	126	50.0	0	50.0	9.54
Gemma-3n-E4B	84	33.3	53	42.2	2.50
Gemma-4-E4B	72	28.6	47	35.1	3.71
Ministral-3-3B	68	27.0	5	27.5	7.99
Qwen3-VL-4B	158	62.7	0	62.7	26.68
Qwen3-VL-8B	182	72.2	0	72.2	39.21
Optimized prompt					
Gemma-3-12B	154	61.1	0	61.1	12.96
Gemma-3n-E4B	97	38.5	0	38.5	3.28
Gemma-4-E4B	51	20.2	106	34.9	4.62
Ministral-3-3B	85	33.7	0	33.7	6.81
Qwen3-VL-4B	176	69.8	0	69.8	28.43
Qwen3-VL-8B	179	71.0	0	71.0	44.12

(b) Class-wise classification performance

Model	Clamp	Needle holder	Scalpel	Shears	Tweezers
Base prompt					
Gemma-3-12B	4/56 (7.1)	2/53 (3.8)	35/41 (85.4)	34/48 (70.8)	51/54 (94.4)
Gemma-3n-E4B	0/56 (0.0)	0/53 (0.0)	25/41 (61.0)	9/48 (18.8)	50/54 (92.6)
Gemma-4-E4B	38/56 (67.9)	0/53 (0.0)	2/41 (4.9)	15/48 (31.2)	17/54 (31.5)
Ministral-3-3B	0/56 (0.0)	0/53 (0.0)	0/41 (0.0)	46/48 (95.8)	22/54 (40.7)
Qwen3-VL-4B	23/56 (41.1)	0/53 (0.0)	33/41 (80.5)	48/48 (100.0)	54/54 (100.0)
Qwen3-VL-8B	34/56 (60.7)	8/53 (15.1)	40/41 (97.6)	46/48 (95.8)	54/54 (100.0)
Optimized prompt					
Gemma-3-12B	6/56 (10.7)	18/53 (34.0)	37/41 (90.2)	46/48 (95.8)	47/54 (87.0)
Gemma-3n-E4B	0/56 (0.0)	3/53 (5.7)	33/41 (80.5)	13/48 (27.1)	48/54 (88.9)
Gemma-4-E4B	0/56 (0.0)	0/53 (0.0)	8/41 (19.5)	37/48 (77.1)	6/54 (11.1)
Ministral-3-3B	0/56 (0.0)	4/53 (7.5)	24/41 (58.5)	45/48 (93.8)	12/54 (22.2)
Qwen3-VL-4B	33/56 (58.9)	0/53 (0.0)	41/41 (100.0)	48/48 (100.0)	54/54 (100.0)
Qwen3-VL-8B	25/56 (44.6)	12/53 (22.6)	40/41 (97.6)	48/48 (100.0)	54/54 (100.0)

Note. Strict accuracy is computed over all 252 images. Valid accuracy uses only predictions matching an allowed class label. Invalid denotes unmappable outputs. Average latency is in seconds per image. Class-wise entries are correct/total.

D.5 VLM Hinge Classification

Table D.5.1. Overall VLM hinge-state classification results on the test set ($n = 157$: 78 open, 79 closed). Strict accuracy is computed over all samples, whereas valid accuracy is computed only over valid predictions. In the confusion-count columns, the *True open* and *True closed* column groups show how samples from each ground-truth state were predicted as open, closed, or invalid.

	Overall				**True open** ($n = 78$)			**True closed** ($n = 79$)		
Model	**Correct**	**Strict acc.**	**Invalid**	**Valid acc.**	**Open**	**Closed**	**Inv.**	**Open**	**Closed**	**Inv.**
Qwen3-VL-8B	**98/157**	**62.4**	0	**62.4**	73	5	0	54	25	0
Gemma-3-12B	87/157	55.4	0	55.4	77	1	0	69	10	0
Qwen3-VL-4B	82/157	52.2	0	52.2	70	8	0	67	12	0
Ministral-3-3B	81/157	51.6	0	51.6	78	0	0	76	3	0
Gemma-3n-E4B	71/157	45.2	0	45.2	29	49	0	37	42	0
Gemma-4-E4B	6/157	3.8	146	54.6	6	0	72	5	0	74

Table D.5.2. Per-instrument VLM hinge-state classification accuracy on the test set. Accuracies are computed over all samples for each instrument, with invalid predictions counted as incorrect. Results are reported for clamps (CM, $n = 56$), needle holders (NH, $n = 53$), and shears (SH, $n = 48$).

Model	**Clamp acc. (%)**	**Needle Holder acc. (%)**	**Shear acc. (%)**
Qwen3-VL-8B	**76.8**	54.7	54.2
Gemma-3-12B	55.4	56.6	54.2
Qwen3-VL-4B	53.6	52.8	50.0
Ministral-3-3B	53.6	**58.5**	41.7
Gemma-3n-E4B	51.8	28.3	**56.2**
Gemma-4-E4B	5.4	5.7	0.0