LD-RSVIS: A LARGE-SCALE AND DIVERSE BENCHMARK FOR REFERRING SURGICAL VIDEO INSTRUMENT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring surgical video instrument segmentation (RSVIS) aims at segmenting the instrument in a surgical video, given a textual description. Despite recent progress, current models are trained and assessed on relatively small-scale benchmarks, hindering the development of more general RSVIS. In addition, existing benchmarks support only the single-target expression that refers to one instrument in the video, while overlooking multi-target and no-target referring expressions, restricting the applicability of RSVIS in practical scenarios. Addressing these issues, we propose ***LD-RSVIS***, a new benchmark aiming to facilitate more robust and general RSVIS. Specifically, LD-RSVIS consists of 3,536 surgical videos with 1.09 million frames and covers a broad set of 30 instrument classes from 25 various procedures. By including abundant videos and classes, LD-RSVIS could benefit large-scale training and evaluation of more general RSVIS methods. Besides, unlike existing datasets, LD-RSVIS offers diverse referring settings, including no-target, single-target, and multi-target expressions, which enables the development of more practical RSVIS models in real applications. In order to ensure high-quality annotations, all videos in LD-RSVIS are manually labeled with multiple rounds of inspection and refinement. To our knowledge, LD-RSVIS is the largest and most diverse benchmark for RSVIS. To analyze LD-RSVIS and to provide comparison for future research, we evaluate 12 representative methods, and the results reveal that more efforts are required for improvements. To encourage future research, we present a simple yet effective RSVIS method, dubbed ***Cascade-RSVIS***, that first mines target-specific cues using the complementary multi-cue text information and then employs such cues and textual information for segmentation, achieving promising performance. Through LD-RSVIS and Cascade-RSVIS, we hope to inspire more future research toward robust and general RSVIS. Our benchmark and code will be released.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Referring surgical video instrument segmentation (RSVIS) aims at segmenting an instrument of interest within a surgical sequence, given the textual expression. Owing to its important applications in various scenarios, including surgical robotics and intelligent surgical assistants, RSVIS has received increasing interest in recent years (Zhou et al., 2023; Wang et al., 2024; Liu et al., 2025; Wei et al., 2026). In particular, several datasets have been introduced (Wang et al., 2024; Liu et al., 2025; Wei et al., 2026; Liu et al., 2026) for training and assessing RSVIS models, greatly facilitating RSVIS. Despite this, these benchmarks suffer from several critical limitations that hinder further development of RSVIS.

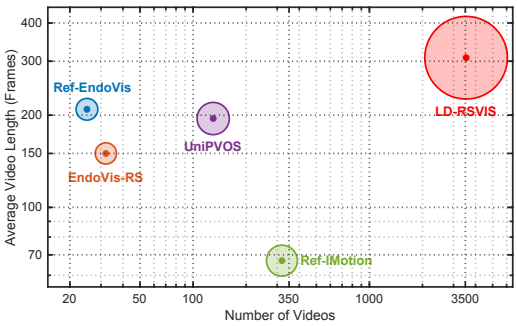


Figure 1: Comparison of existing RSVIS benchmarks. LD-RSVIS is larger than all other benchmarks. *Best viewed in color for all figures.*

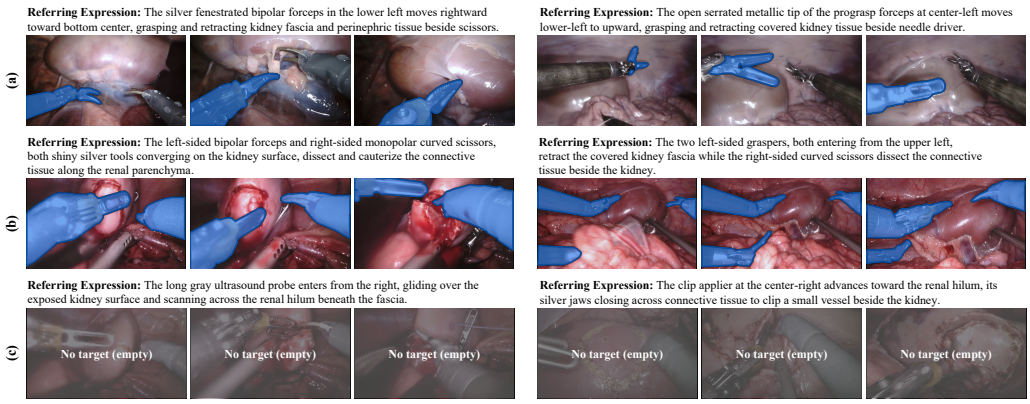


Figure 2: Illustration of different referring settings in RSVIS. Image (a): Single-target expression in LD-RSVIS and other benchmarks; Image (b): Multi-target expression in our LD-RSVIS; Image (c): No-target expression in LD-RSVIS. All examples are from LD-RSVIS.

One of the major issues with existing RSVIS benchmarks is their relatively small scale. Particularly, in the deep learning era, abundant surgical sequences are crucial for learning robust RSVIS models. However, as displayed in Fig. 1, all current RSVIS datasets contain *fewer than* 350 video sequences, making them *inadequate* for large-scale model training and thereby hindering the further advances in RSVIS. Beyond facilitating model training, a large-scale benchmark is also essential for reliable and comprehensive evaluation. Nevertheless, existing RSVIS datasets comprise *less than* 100 sequences in their test sets, making the evaluation susceptible to bias and potentially sensitive to a small number of individual videos. Moreover, such limited test sets in current benchmarks may fail to adequately capture diverse scenarios in the real-world surgeries. Consequently, the reported performance may not faithfully reflect the true capability and generality of RSVIS methods.

Besides the small-scale problem, another limitation of current benchmarks is the restricted referring setting. Specifically, existing benchmarks only support the single-target expression (see Fig. 2 (a)), which restricts RSVIS models to predicting one and only one target. In practical scenarios, however, a referring expression may simultaneously describe multiple instruments (see Fig. 2 (b)), particularly when they are involved in the same surgical interaction, requiring the model to identify and segment all matching targets. Furthermore, in certain situations, the target of interest may not appear in the video sequence at all (see Fig. 2 (c)), requiring the model to recognize the absence of the referred target, instead of incorrectly matching the expression to an irrelevant target. To better reflect diverse referring needs in real-world surgeries, a benchmark should support more comprehensive referring settings, including no-target, single-target, and multi-target expressions.

Contribution. To alleviate the above limitations and to further facilitate RSVIS, we present a Large-scale and Diverse benchmark for RSVIS, named ***LD-RSVIS***. Specifically, LD-RSVIS contains 3,536 surgical video sequences with 1.09 million frames, and covers a broad set of 30 instrument categories from 25 various procedures. Compared to current benchmarks, our LD-RSVIS is 10$\times$ larger in scale, enabling more robust training and more faithful evaluation of RSVIS approaches. In addition, unlike existing datasets that support only the single-target expressions, LD-RSVIS, inspired by generalized referring image segmentation (Liu et al., 2023), provides more diverse referring settings, including no-target expressions, where the target of interest is absent; single-target expressions, where the text refers to a single target; and multi-target expressions, where text refers to multiple target objects. In total, LD-RSVIS provides 58,666 textual expressions across these three referring settings, aiming at facilitating the development of more practical RSVIS systems. To ensure the high annotation quality, all videos in LD-RSVIS are manually labeled and undergo multiple rounds of careful inspection and refinement. To our best knowledge, LD-RSVIS is currently the *largest* RSVIS benchmark, and also the *first* to introduce various referring settings into RSVIS.

In order to understand how existing method perform and to provide comparisons for future research, we extensively evaluate 12 representative methods that provide executable implementations, including recent RSVIS approaches and general referring video object segmentation (RVOS) models, and conduct in-depth comparison and analysis. Our evaluation results (as described later) display that, not surprisingly, these top-performing approaches perform substantially below satisfactory levels

Table 1: Comparison of our LD-RSVIS and current RSVIS datasets. "NT", "ST", and "MT" indicate no-target, single-target, and multi-target, respectively. "n/a" denotes that data is not available. The video frame rate of all datasets is 1 frame per second (fps).

Benchmark	Year	Videos	Total Frames	Instru. Classes	Avg. Vid. Length	Num. of Expressions	Num. of Masks	Referring Expression		
								ST	MT	NT
EndoVis-RS (Wang et al., 2024)	2024	32	4.8**K**	9	150	n/a	9.8**K**	✓	✗	✗
Ref-EndoVis (Liu et al., 2025)	2025	25	5.2**K**	9	209	n/a	12.3**K**	✓	✗	✗
UniPVOS (Liu et al., 2026)	2026	130	25.0**K**	n/a	195	1,367	178.8**K**	✓	✗	✗
Surg-IMotion (Wei et al., 2026)	2026	319	21.0**K**	n/a	67	718	n/a	✓	✗	✗
LD-RSVIS (Ours)	2026	3,536	1.09**M**	30	308	58,666	3.85**M**	✓	✓	✓

on the more challenging LD-RSVIS benchmark. This demonstrates the difficulties in achieving robust and generic referring surgical video instrument segmentation in real-world scenarios, and more efforts are required to improve the performance for practical applications.

Moreover, to facilitate future research, we further introduce a simple yet effective method for RSVIS, dubbed ***Cascade-RSVIS***. Specifically, considering that the given expression alone is *insufficient* to describe the targets in a video involving severe appearance variations or distractors, Cascade-RSVIS, built on SAM 2 (Ravi et al., 2025), employs a two-stage strategy, consisting of target-context mining (TCM) and context-guided segmentation (CGS), for RSVIS. More concretely, TCM aims to generate target-specific appearance cues in a video using textual information encoded in referring expression, while CGS leverages such target appearance cues to enhance the memory and further incorporates the textual feature for more accurate target segmentation. By connecting TCM and CGS in a unified framework, Cascade-RSVIS is able to benefit from additional discriminative target appearance cues, enabling more robust segmentation. In addition, to better exploit target-relevant information from the expression, we present a multi-cue text representation that combines the global textual semantics with learned complementary semantic cues, and applies it in both TCM and CGS, further improving the performance. Despite simplicity, Cascade-RSVIS achieves promising results on LD-RSVIS, and is expected to provide a solid reference for future research.

In summary, our main contributions are as follows: ♠ We introduce LD-RSVIS, a large-scale dataset comprising 3,536 surgical videos with 1.09 million frames and diverse referring settings for RSVIS; ♥ We evaluate 12 representative methods to understand their performance and to offer comparisons to future research; ♣ We propose a simple yet effective method, Cascade-RSVIS, to facilitate more future research on RSVIS; ♦ Cascade-RSVIS shows promising results on LD-RSVIS, establishing a solid reference and providing guidance for future research.

2 Related Work

RSVIS Benchmarks. Benchmarks play an important role for the development of RSVIS. EndoVis-RS (Wang et al., 2024) is the first benchmark for referring surgical video instrument segmentation. It is constructed from the popular EndoVis challenges (Allan et al., 2019; 2020) and consists of 32 sequences with 4.8K frames. Ref-EndoVis (Liu et al., 2025) is built upon EndoVis-RS (Wang et al., 2024) with refined annotations by addressing inconsistencies and omissions in the original masks, comprising 25 videos and 5.2K frames. Later in (Liu et al., 2026), Ref-EndoVis is further extended into a unified dataset UniPVOS by including extra video sequences. Notably, both Ref-EndoVis and UniPVOS extend the scope of referring surgical video segmentation beyond surgical instruments by additionally incorporating tissues as referred targets. More recently, Ref-IMotion (Wei et al., 2026) offers 319 video sequences with 21K frames to facilitate motion-aware RSVIS. Despite these efforts, existing datasets remain limited in scale, hindering further advances in RSVIS. To alleviate this, LD-RSVIS provides 3,536 videos with 1.09M frames, which is 10× larger than current datasets. In addition, unlike existing benchmarks supporting only single-target expressions, LD-RSVIS offer more diverse referring settings, including no-target, single-target, and multi-target expressions, for more practical RSVIS. Tab. 1 summarizes the differences between LD-RSVIS and existing datasets.

RSVIS Algorithms. RSVIS has recently drawn growing interest owing to its important applications in surgical robotics and intelligent surgical assistants. The method of (Zhou et al., 2023) introduces a text-promptable architecture for surgical instrument segmentation by leveraging text-based category prompts and multimodal vision-language models. The work of (Wang et al., 2024) proposes a video-instrument synergistic learning method for text-referred instrument segmentation. The work

of (Liu et al., 2025) introduces the credible initial-frame selection and diversity-driven memory to improve SAM 2 (Ravi et al., 2025) for RSVIS, which is further extended in (Liu et al., 2026) by unifying multiple prompts. The method of (Wei et al., 2026) presents a motion-aware framework for RSVIS. Different from these methods, Cascade-RSVIS introduces a two-stage architecture that first mines target-specific cues from the video and then employs such cues for segmentation, enabling the exploration of additional discriminative target appearance cues for robust performance.

Referring Video Object Segmentation. Referring video object segmentation (RVOS) aims to segment the target given a natural language description, and has been extensively studied over the past decade with many datasets (Gavrilyuk et al., 2018; Khoreva et al., 2018; Seo et al., 2020; Ding et al., 2023; Liang et al., 2026) and methods (Wu et al., 2022; 2023; Luo et al., 2023; Liang et al., 2025; Pan et al., 2025; Cheng et al., 2026; Jiang et al., 2026). Different from these works, we specifically focus on segmenting referred instruments in surgical videos, where frequent occlusions, visually similar distractors, and repeated target entry and exit pose substantial difficulties for accurate target identification and segmentation. To advance research in this direction, we present a new benchmark for referring surgical instrument segmentation, together with an effective baseline method.

3 The Proposed LD-RSVIS Benchmark

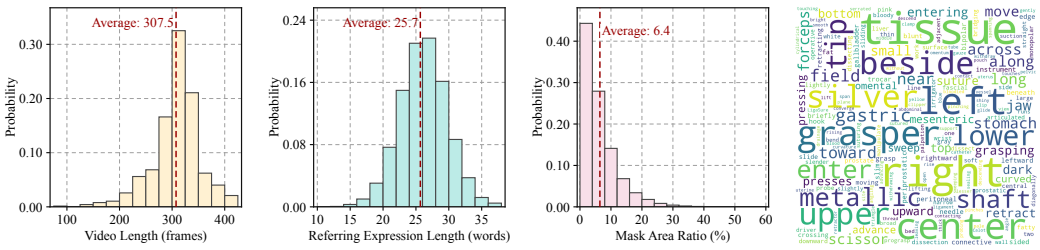
3.1 Construction Principle

LD-RSVIS aims to offer a dedicated dataset to foster the development of RSVIS. For this purpose, we follow several principles in constructing LD-RSVIS: **(i)** ***Large-scale.*** Large-scale benchmark is crucial for training robust RSVIS models and for providing comprehensive evaluation. Therefore, we expect LD-RSVIS to contain at least 3,000 videos, benefiting both large-scale model training and assessment. **(ii)** ***Diverse referring settings.*** In real-world surgical scenarios, the referring expression may describe one or several instruments or have no matching target. To this end, LD-RSVIS aims to offer more diverse referring settings, including no-target, single-target, and multi-target expression, enable more practical RSVIS. **(iii)** ***Rich procedures and instrument categories.*** An important goal of LD-RSVIS is to advance the development of general RSVIS which aims at segmenting instruments of different classes in various procedures. To this end, the new benchmark is designed to cover a broad range of instrument categories and surgical procedures. **(iv)** ***High quality.*** High-quality annotations are essential for both training and evaluation. To ensure the high quality of LD-RSVIS, all sequences are manually labeled and undergo multiple rounds of careful inspection and refinement.

3.2 Data Acquisition

Similar to existing RSVIS benchmarks (Wang et al., 2024; Liu et al., 2025; 2026; Wei et al., 2026), we construct LD-RSVIS by collecting publicly available videos from current surgical video datasets. However, unlike prior RSVIS datasets that primarily gather videos from surgical video segmentation datasets, we broaden the source pool to contain a wider range of surgical video understanding tasks, thereby increasing the diversity of surgical instruments and procedures in our dataset. Specifically, we source videos from datasets developed for surgical scene segmentation and/or detection (Allan et al., 2019; 2020; Murali et al., 2023; Ayobi et al., 2025; Wang et al., 2022; Maier-Hein et al., 2021; Yoon et al., 2021), surgical phase and action recognition (Rueckert et al., 2026; Lavanchy et al., 2024; Bawa et al., 2021; Psychogyios et al., 2024), surgical workflow understanding (Twinanda et al., 2017; Wagner et al., 2023), surgical task and activity recognition (Zia et al., 2025), and large-scale surgical video pre-training (Jaspers et al., 2026). Together, these diverse sources offer a broad basis for constructing a more general large-scale and diverse RSVIS benchmark. Due to space limitation, we present the data selected from each source in the ***appendix***.

Once the source pool is determined, we conduct a verification process in which the videos from each source dataset are carefully inspected to assess their suitability for the RSVIS task. This process is carried out by several experts (students who work on related surgical video understanding tasks) to ensure the quality of selected videos. Specifically, for each sequence, we first remove the corrupted or irrelevant content. If the remaining video comprises at least one clip suitable for RSVIS, we retain this sequence; otherwise, it is discarded. Since many of the original surgical videos are considerably long, often spanning tens of minutes to several hours, we follow a similar strategy to existing RSVIS benchmarks and sample several representative clips rather than using the entire video. These selected



(a) Distribution of video length (b) Distribution of expression length (c) Distribution of mask area ratio (d) Word cloud of referring expressions

Figure 3: Representative statistics of LD-RSVIS, comprising distributions of video length, referring expression length, mask area ratio, and wordcloud of referring expressions.

clips are chosen to capture diverse instrument appearances and classes, surgical stages, and temporal conditions, reducing temporal redundancy while preserving diversity of the original surgical videos. Due to space limitation, we show statistics of selected clips from each video source in the ***appendix***.

Eventually, we compile a large-scale benchmark, LD-RSVIS, for RSVIS. LD-RSVIS contains 3,536 videos with 1.09 million frames and covers 30 instrument classes from 25 procedures. Each video is paired with multiple referring expressions covering different referring scenarios, including no-target, single-target, and multi-target cases, resulting in a total of 58,666 expressions. Compared to current RSVIS datasets, LD-RSVIS is much larger and more diverse. Tab. 1 summarizes LD-RSVIS and existing benchmarks. Due to limited space, additional details on the instrument classes and surgical procedures, as well as discussions of dataset maintenance, ethical considerations, and responsible usage of the proposed LD-RSVIS, are provided in the ***appendix***.

3.3 Annotation

After data collection, our experts first generate referring expressions for each sequence. The referring expressions can describe instruments at different levels of granularity, covering both complete instruments and their semantic parts, enabling more fine-grained and flexible referring segmentation. Different from current RSVIS benchmarks, we, inspired by generalized referring image segmentation (Liu et al., 2023), provide different types of expressions, including no-target, single-target, and multi-target referring expressions. In a no-target expression, the referred target is absent; in a single-target expression, the text refers to a single target; in a multi-target expression, the text refers to a varying number (larger than one) of target objects. It is worth noting that, for videos containing only one valid target, multi-target expressions are not provided. In total, 58,666 referring expressions are generated in LD-RSVIS. Afterwards, we label the corresponding targets referred in the expressions from videos with consistent spatio-temporal masklets.

To ensure high-quality annotations of LD-RSVIS, we compile an annotation team, composed of an expert group (students working on related surgical video understanding tasks) for verification and a volunteer group for labeling, and use a multi-round annotation strategy. In specific, the expert group first label the referred targets in initial frames. Then, the volunteer group, divided into several teams, works to annotate the complete masklet based on the initial mask for each referring expression. After this initial round, the masklet annotations are sent to the expert group for inspection, where the mask completeness and boundaries, temporal consistency, and target identity will be carefully reviewed. If the initial annotations are not unanimously agreed by all experts, they will be returned back to the original labeling team for refinement. We repeat this process until all annotations satisfy the quality requirements. Due to space limitation, we show the pipeline of annotation process in the ***appendix***. Fig. 2 shows several annotation examples in LD-RSVIS, and more can be seen in the ***appendix***.

Statistics of annotation. To better understand LD-RSVIS, we demonstrate representative statistics in Fig. 3, including the distributions of the video length, referring expression length, mask area ratio (defined as the percentage of pixels occupied by referred targets in the corresponding frame), and wordcloud of referring expressions. As displayed in Fig. 3, LD-RSVIS contains videos and referring expressions of varying lengths and target objects of different scales. Together, these statistics exhibit the diversity of LD-RSVIS, making it a challenging and realistic benchmark for RSVIS.

3.4 Dataset Split and Evaluation Metric

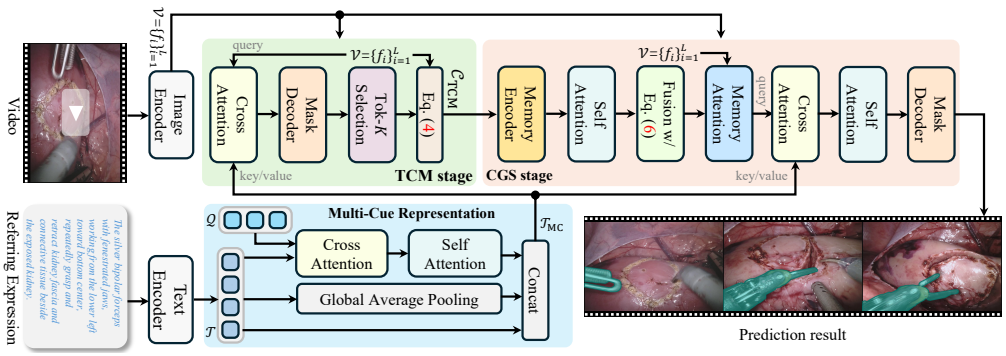


Figure 4: Overview of Cascade-RSVIS that first mines target-relevant context via TCM and then leverages such context together with the referring expression for segmentation through CGS.

Dataset Split. LD-RSVIS contains 3,536 videos, with 2,853 sequences selected for the training set, named LD-RSVIS$_{\text{Tra}}$, and the rest 683 for test set, dubbed LD-RSVIS$_{\text{Tst}}$. To prevent train-test leakage, video clips derived from the same original long surgical video are assigned exclusively to either training set or test set. Besides, both LD-RSVIS$_{\text{Tra}}$ and LD-RSVIS$_{\text{Tst}}$ follow the same referring settings, comprising no/single/multi-target expressions. Tab. 2 summarizes training and test sets.

Table 2: Comparison of training and test sets.

	Videos	Total Frames	Avg. Length	Num. of Expre.	Num. of Masks
LD-RSVIS$_{\text{Tra}}$	2,853	880,080	308	47,500	3,135,866
LD-RSVIS$_{\text{Tst}}$	683	207,282	303	11,166	711,398

Evaluation Metrics. Following (Liu et al., 2025), we apply two metrics for evaluation, including region similarity $\mathcal{J}$, which computes the Intersection over Union (IoU) of the predicted and ground-truth masks, and contour accuracy $\mathcal{F}$ (Perazzi et al., 2016), which evaluates the contour accuracy of prediction results. For comprehensive evaluation, we further calculate the average of $\mathcal{J}$ and $\mathcal{F}$, denoted as $\mathcal{J}\&\mathcal{F}$. It is worth noting that, following (Ding et al., 2025), for a no-target expression, a prediction is considered correct only if the output remains empty throughout the entire sequence. In this case, both $\mathcal{J}$ and $\mathcal{F}$ are set to 1; otherwise, both are set to 0. Due to limited space, please refer to (Perazzi et al., 2016; Ding et al., 2025; Liu et al., 2025) for more details of these metrics.

4 The Proposed Cascade-RSVIS

Overview. We propose Cascade-RSVIS for RSVIS by extending SAM 2 (Ravi et al., 2025), which exploits target-specific appearance cues from a video through a two-stage framework. As in Fig. 4, after feature extraction, the target-context mining (TCM) stage first mines reliable target context using the referring expression, and the context-guided segmentation (CGS) stage then leverages such context together with the referring expression for segmentation.

4.1 Feature Extraction

In Cascade-RSVIS, we first extract visual and textual features for the video and referring expression.

Video Feature Extraction. Given a video with L frames $\mathcal{I} = \{I_i\}_{i=1}^{L}$, we apply the image encoder $\Phi(\cdot)$ from SAM 2 (Ravi et al., 2025) to extract visual features $\mathcal{V}$ via $\mathcal{V} = \{f_i\}_{i=1}^{L}$, where $f_i = \Phi(I_i)$. The visual features will be used in the subsequent stages for target context mining and segmentation.

Multi-Cue Textual Feature Extraction for Expression. Referring expressions often contain multiple complementary cues to identify the target, which may not be fully captured by a single textual representation. Hence, we construct a multi-cue textual feature that combines word-level, holistic, and learned complementary semantics. Specifically, as in Fig. 4, we first tokenize the expression into $\mathcal{W} = \{w_j\}_{j=1}^{N_t}$ and employ RoBERTa (Liu et al., 2019) to obtain word embeddings $\mathcal{T} = \{t_j\}_{j=1}^{N_t}$. We then perform average pooling over $\mathcal{T}$ to obtain a holistic sentence feature $\tilde{\mathcal{T}} = \texttt{AvgPool}(\mathcal{T})$, which summarizes overall semantics of the expression. To further capture fine-grained complementary cues in the expression, we introduce N_q learnable queries $\mathcal{Q} = \{q_n\}_{n=1}^{N_q}$. These queries retrieve information from the word sequence through cross-attention and subsequently interact via self-

attention, yielding $\mathcal{U} = \texttt{SA}(\texttt{CA}(\mathcal{Q}, \mathcal{T}))$, where $\texttt{SA}(\mathbf{z})$ denotes the self-attention block with $\mathbf{z}$ generating query/key/value, and $\texttt{CA}(\mathbf{z}, \mathbf{u})$ the cross-attention block with $\mathbf{z}$ generating query and $\mathbf{u}$ key/value, as in (Vaswani et al., 2017). Finally, we concatenate the word feature $\mathcal{T}$, sentence feature $\tilde{\mathcal{T}}$, and query features $\mathcal{U}$ to produce the final multi-cue textual representation $\mathcal{T}_{\text{MC}} = \texttt{Concat}(\mathcal{T}, \tilde{\mathcal{T}}, \mathcal{U})$.

4.2 Our Cascade-RSVIS Framework

After extracting video and expression features, Cascade-RSVIS performs segmentation through two cascaded stages, including TCM and CGS, as illustrated in the following.

Target-Context Mining (TCM). The TCM stage mines target-relevant context from the video based on the referring expression. To this end, we first inject the textual feature into each frame through cross-attention followed by a residual connection to obtain the multimodal video features $\mathcal{V}_{\text{TCM}}$ via

$$\mathcal{V}_{\text{TCM}} = \{f_i^{\text{TCM}}\}_{i=1}^{L} \qquad f_i^{\text{TCM}} = \texttt{CA}(f_i, \mathcal{T}_{\text{MC}}) + f_i \tag{1}$$

where $\texttt{CA}(\cdot,\cdot)$ is the cross-attention block. Afterwards, we apply mask decoder $\texttt{MaskDec}(\cdot)$ in (Ravi et al., 2025) to each f_i^{TCM} of $\mathcal{V}_{\text{TCM}}$, as follows,

$$M_i^{\text{TCM}} = \texttt{MaskDec}(f_i^{\text{TCM}}) \tag{2}$$

where M_i^{TCM} denotes the prediction in frame i, and contains a mask m_i^{TCM}, an IoU score s_i^{TCM}, and an object-presence score p_i^{TCM}.

Given the predicted results $R^{\text{TCM}} = \{M_i^{\text{TCM}}\}_{i=1}^{L}$ over the video, we select target context from these predictions. Following SAM 2 (Ravi et al., 2025), if the maximum object-presence score of R^{TCM} is below zero, we directly return an empty mask sequence. Otherwise, we retain predictions with non-negative object-presence scores and select the top K indices according to their predicted IoU scores, as follows,

$$\mathcal{K} = \texttt{IdxTopK}_{i\in\mathcal{D}}(s_i^{\text{TCM}}) \qquad \mathcal{D} = \{i | p_i^{\text{TCM}} \geq 0\} \tag{3}$$

where $\mathcal{K}$ denotes indices of the selected top K predictions, and $\texttt{IdxTopK}(\cdot)$ returns indices of the top K predictions ranked by IoU scores. Based on $\mathcal{K}$, the target context $\mathcal{C}_{\text{TCM}}$ is formed as follows,

$$\mathcal{C}_{\text{TCM}} = \{(f_i, m_i^{\text{TCM}}, s_i^{\text{TCM}})\}_{i\in\mathcal{K}} \tag{4}$$

The resulting $\mathcal{C}_{\text{TCM}}$ is then passed to the subsequent CGS stage for segmentation.

Context-Guided Segmentation (CGS). The CGS stage aims to leverage the context $\mathcal{C}_{\text{TCM}}$ from TCM to guide segmentation. Specifically, for each selected context in $\mathcal{C}_{\text{TCM}}$, we first generate its memory by encoding its video feature and mask with memory encoder $\texttt{MemEnc}(\cdot,\cdot)$ in SAM 2 via

$$\mathcal{M}_i = \texttt{MemEnc}(f_i, m_i^{\text{TCM}}) \quad i \in \mathcal{K} \tag{5}$$

Once obtaining $\mathcal{M}_{\text{CGS}} = \{\mathcal{M}_i\}_{i\in\mathcal{K}}$, we perform self-attention on it for feature interaction via $\bar{\mathcal{M}}_{\text{CGS}} = \{\bar{\mathcal{M}}_i\}_{i\in\mathcal{K}} = \texttt{SA}(\mathcal{M}_{\text{CGS}})$ with $\bar{\mathcal{M}}_i$ the memory feature in frame i after interaction, and then aggregate memory features in $\bar{\mathcal{M}}_{\text{CGS}}$ based on their normalized IoU scores, as follows,

$$\tilde{\mathcal{M}}_{\text{CGS}} = \sum\nolimits_{i\in\mathcal{K}} \omega_i \cdot \bar{\mathcal{M}}_i \qquad \omega_i = \texttt{exp}(s_i^{\text{TCM}}) / \sum\nolimits_{j\in\mathcal{K}} \texttt{exp}(s_j^{\text{TCM}}) \tag{6}$$

where $\tilde{\mathcal{M}}_{\text{CGS}}$ denotes the fused memory feature derived from the context $\mathcal{C}_{\text{TCM}}$, which is injected into the video features via memory attention $\texttt{MemAtt}(\cdot,\cdot)$ in SAM 2, as follows,

$$\mathcal{V}_{\text{CGS}} = \{f_i^{\text{CGS}}\}_{i=1}^{L} \qquad f_i^{\text{CGS}} = \texttt{MemAtt}(f_i, \tilde{\mathcal{M}}_{\text{CGS}}) \tag{7}$$

With memory-enhanced video features $\mathcal{V}_{\text{CGS}}$, we further inject the expression feature into each f_i^{CGS} to obtain the multimodal feature g_i^{CGS} through cross-attention followed by a residual connection via

$$g_i^{\text{CGS}} = \texttt{CA}(f_i^{\text{CGS}}, \mathcal{T}_{\text{MC}}) + f_i^{\text{CGS}} \tag{8}$$

With $\mathcal{G} = \{g_i^{\text{CGS}}\}_{i=1}^{L}$, we further perform self-attention to model temporal interaction in the video via $\tilde{\mathcal{G}} = \{\tilde{g}_i^{\text{CGS}}\}_{i=1}^{L} = \texttt{SA}(\mathcal{G})$, where $\texttt{SA}(\cdot)$ represents the self-attention block. Afterwards, we apply the mask decoder $\texttt{MaskDec}(\cdot)$ to each $\tilde{g}_i^{\text{CGS}}$ and retain its predicted mask, as follows,

$$m_i^{\text{CGS}} = \texttt{MaskDec}(\tilde{g}_i^{\text{CGS}}) \tag{9}$$

where m_i^{CGS} denotes predicted mask segmentation in each frame i.

Table 3: Comparison with existing methods on LD-RSVIS$_{\text{Tst}}$. "MTE", "STE", and "NTE" indicate evaluations on multi-target, single-target, and no-target referring expressions, respectively.

	Overall			MTE			STE			NTE
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
VIS-Net (Wang et al., 2024)$_{\text{[T-MI'24]}}$	20.8	18.6	19.7	19.7	15.5	17.6	24.4	22.1	23.3	14.7
ReferFormer (Wu et al., 2022)$_{\text{[CVPR'22]}}$	19.0	23.6	21.3	23.3	31.9	27.6	21.2	26.1	23.7	10.3
OnlineRefer (Wu et al., 2023)$_{\text{[ICCV'23]}}$	20.4	24.2	22.3	27.8	34.0	30.9	21.5	26.0	23.7	10.9
LoSh (Yuan et al., 2024)$_{\text{[CVPR'24]}}$	22.5	28.1	25.3	26.4	38.8	32.6	26.0	31.0	28.5	11.7
LMPM (Ding et al., 2023)$_{\text{[ICCV'23]}}$	25.5	27.3	26.4	37.4	39.0	38.2	21.0	23.8	22.4	22.7
ReSurgSAM2 (Liu et al., 2025)$_{\text{[MICCAI'25]}}$	27.0	29.6	28.3	29.1	33.1	31.1	26.4	29.6	28.0	26.1
SurgRef (Wei et al., 2026)$_{\text{[AAAI'26]}}$	29.4	27.8	28.6	38.2	36.0	37.1	27.8	25.7	26.8	23.8
DsHmp (He & Ding, 2024)$_{\text{[CVPR'24]}}$	24.7	33.3	29.0	28.1	46.9	37.5	22.9	30.7	26.8	25.0
SAMWISE (Cuttano et al., 2025)$_{\text{[CVPR'25]}}$	28.8	32.2	30.5	32.2	38.0	35.1	33.8	37.7	35.8	15.4
ReferDINO (Liang et al., 2025)$_{\text{[ICCV'25]}}$	29.2	33.6	31.4	30.1	37.5	33.8	34.7	39.8	37.3	17.3
ReferMo (Liang et al., 2026)$_{\text{[CVPR'26]}}$	31.6	32.0	31.8	40.6	36.4	38.5	36.7	39.6	38.2	12.4
FlowRVS (Wang et al., 2026)$_{\text{[ICLR'26]}}$	29.8	36.4	33.1	31.1	46.3	38.7	36.0	41.6	38.8	16.2
Cascade-RSVIS (ours)	36.1	38.7	37.4	37.8	41.8	39.8	38.0	41.2	39.6	30.6

Optimization. In Cascade-RSVIS, the segmentation predictions from both TCM and CGS are optimized during training. Due to space limitations, the loss functions are provided in the ***appendix***.

5 Experiments

Implementation of Cascade-RSVIS. Cascade-RSVIS is implemented in PyTorch (Paszke et al., 2019) and trained on two NVIDIA A100 (40G) GPUs. The image encoder $\Phi(\cdot)$ adopts the Hiera-B+ (Ryali et al., 2023). Cascade-RSVIS is trained in two stages: TCM for 18 epochs and CGS for 24 epochs, with a batch size of 2 using AdamW (Loshchilov & Hutter, 2019). We first optimize TCM and then freeze it while training CGS. The initial learning rates are set to 5×10^{-6} for pretrained parameters and 1×10^{-4} for newly introduced modules. During inference, the clip length L is set to 8 due to memory constraints, and predictions from all clips are combined to obtain the final video segmentation results. The parameters N_q and K are both empirically set to 3.

Evaluated Methods. We evaluate 12 representative approaches with publicly available implementations on LD-RSVIS. These contain three surgical RSVIS methods, including VIS-Net (Wang et al., 2024), ReSurgSAM2 (Liu et al., 2025), and SurgRef (Wei et al., 2026), and nine general RVOS models, including ReferFormer (Wu et al., 2022), OnlineRefer (Wu et al., 2023), LMPM (Ding et al., 2023), LoSh (Yuan et al., 2024), DsHmp (He & Ding, 2024), SAMWISE (Cuttano et al., 2025), ReferDINO (Liang et al., 2025), ReferMo (Liang et al., 2026), and FlowRVS (Wang et al., 2026). It is worth noting that, all the evaluated methods are trained on LD-RSVIS$_{\text{Tra}}$ for fair comparison.

5.1 Evaluation Results

Overall evaluation. We demonstrate results of evaluated approaches and Cascade-RSVIS in Tab. 3. Our Cascade-RSVIS achieves the best overall $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$ scores of 36.1%, 38.7%, and 37.4%, respectively. Specifically, in terms of $\mathcal{J}\&\mathcal{F}$, it outperforms the second best method, FlowRVS with 33.1% $\mathcal{J}\&\mathcal{F}$ score, by 4.3%. Cascade-RSVIS also improves the overall $\mathcal{J}$ score of ReferMo by 4.5% and the overall $\mathcal{F}$ score of FlowRVS by 2.3%. These results consistently validate the efficacy of our Cascade-RSVIS by mining and leveraging target context from videos for segmentation.

Evaluation across different expressions. To enable in-depth analysis of methods, we show results across multi-target (2,732 expressions), single-target (6,573 expressions), and no-target referring expressions (1,861 expressions) in Tab. 3. Please note that, for no-target expressions, $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$ are identical because both $\mathcal{J}$ and $\mathcal{F}$ are set to 1 for an entirely empty prediction and 0 otherwise; therefore, we report only $\mathcal{J}\&\mathcal{F}$. As shown in Tab. 3, Cascade-RSVIS shows the best $\mathcal{J}\&\mathcal{F}$ scores of 39.8%, 41.2%, and 30.6% on multi-, single-, and no-target expressions, showing its effectiveness.

5.2 Ablation Study on Cascade-RSVIS

We conduct ablations for Cascade-RSVIS on LD-RSVIS$_{\text{Tst}}$. Our final setting is highlighted in cyan.

Table 4: Ablation on the cascade framework.

	Framework	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
❶	TCM alone	28.9	32.5	30.7
❷	Cascade TCM and CGS	36.1	38.7	37.4

Table 5: Ablation on K in TCM.

	Top K predictions	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
❶	$K = 1$	34.9	37.5	36.2
❷	$K = 3$	36.1	38.7	37.4
❸	$K = 5$	35.9	38.3	37.1

Table 6: Ablation on multi-cue representation.

	Word feat.	Sentence feat.	Query feat.	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
❶	✓	-	-	33.7	36.3	35.0
❷	✓	✓	-	34.6	37.6	36.1
❸	✓	-	✓	35.2	38.0	36.6
❹	✓	✓	✓	36.1	38.7	37.4

Table 7: Ablation on context fusion in CGS.

	Context fusion strategy	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
❶	Addition	34.8	37.2	36.0
❷	Gated fusion	35.6	38.2	36.9
❸	IoU-weighted addition	36.1	38.7	37.4

Ablation on cascade framework. Cascade-RSVIS connects TCM and CGS stages in a cascaded manner, where TCM mines target context and CGS exploits it for segmentation. As TCM can directly produce frame-level masks, we compare its standalone predictions with the complete cascade in Tab. 4. As in Tab. 4, TCM alone achieves 30.7% $\mathcal{J}\&\mathcal{F}$ score (❶). When cascading TCM and CGS, the $\mathcal{J}\&\mathcal{F}$ score is significantly improved to 37.4% with by 6.7% (❷) owing to the exploration of context as a guidance for segmentation. This substantial improvement validates the effectiveness of the proposed cascaded framework.

Ablation on K in TCM. We select the top K predictions in TCM to obtain the target context $\mathcal{C}_{\text{TCM}}$. We conduct an ablation on K in Tab. 5. As shown in Tab. 5, $K = 3$ achieves the best $\mathcal{J}\&\mathcal{F}$ of 37.4% (❷), while a larger $K = 5$ slightly degrades the performance (❸), likely due to less reliable target cues and background noise. We therefore set K to 3.

Ablation on multi-cue representation for expression. We construct a multi-cue representation to extract expression feature by combining word feature, global sentence feature, and learned query features. To analyze this, we conduct an ablation in Tab. 6. As in Tab. 6, when using only the word feature alone, the $\mathcal{J}\&\mathcal{F}$ score is 35.0% (❶). Adding the sentence feature improves $\mathcal{J}\&\mathcal{F}$ from 35.0% to 36.1% (❷), while introducing the learned query feature improves it from 35.0% to 36.6% (❸), both enhancing the performance. When combining all features for expression representation, we achieve the best $\mathcal{J}\&\mathcal{F}$ score of 37.4% (❹), validating the efficacy of our multi-cue representation for RSVIS.

Ablation on context fusion in CGS. We aggregate context information in CGS using normalized IoU scores to obtain the fused memory. We compare three fusion strategies, including simple addition with equal context weights, gated fusion with learned context weights, and our IoU-weighted fusion in Tab. 7. As shown, our strategy achieves the best $\mathcal{J}\&\mathcal{F}$ score of 37.4% (❸). This result shows that the mask-quality scores provide a reliable signal for context aggregation.

Table 8: Ablation on N_q.

	Num. of queries N_q	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
❶	$N_q = 1$	34.9	38.1	36.5
❷	$N_q = 3$	36.1	38.7	37.4
❸	$N_q = 5$	35.5	38.3	36.9
❹	$N_q = 7$	35.3	38.1	36.7

Ablation on N_q in the multi-cue representation. In the multi-cue representation, we introduce N_q learnable queries to capture fine-grained complementary cues in the expression. To study the impact of N_q, we conduct an ablation in Tab. 8. As shown in Tab. 8, using 3 learnable queries achieves the best $\mathcal{J}\&\mathcal{F}$ score of 37.4% (❷), while using fewer (❶) or more (❸ and ❹) queries degrades performance, likely due to insufficient semantic coverage with smaller N_q or redundant and noisy cues with larger N_q, respectively. Thus, we set N_q to 3.

Due to limited space, additional details, results, and analyses are provided in the ***appendix***.

6 Conclusion

In this work, we present LD-RSVIS, a large-scale and diverse RSVIS benchmark. LD-RSVIS provides 3,536 videos with 1.09 million frames and diverse referring settings. To our best knowledge, LD-RSVIS is the largest benchmark for RSVIS and the first to support diverse referring settings. Moreover, to encourage future research, we introduce Cascade-RSVIS, a simple yet effective method for RSVIS. Our results demonstrate the advantages of our Cascade-RSVIS over existing methods.

AI USE STATEMENT

In this work, the generative AI tools are used *solely* for language polishing and readability improvement. We do *not* use the generative AI to generate datasets, produce experimental results, formulate mathematical claims, design experiments, implement methods, analyze results, or conduct the final interpretation of the findings. All AI-assisted text has been carefully reviewed, revised, and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

The construction of LD-RSVIS strictly follows ethical standards. LD-RSVIS is constructed from 15 publicly available surgical video sources, and can be used for research purpose only. All the sequences in LD-RSVIS remain subject to licenses, data-use agreements, and attribution requirements of their respective sources. Nevertheless, we understand that the availability or license of certain source videos might change in the future. Once any notification regarding this is received, we will take appropriate actions to handle it.

REFERENCES

Max Allan, Alexey Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis C. García-Peraza, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 2017 robotic instrument segmentation challenge. *arXiv*, 2019.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, Avinash Kori, Varghese Alex, Ganapathy Krishnamurthi, David Rauber, Robert Mendel, Christoph Palm, Sophia Bano, Guinther Saibro, Chi-Sheng Shih, Hsun-An Chiang, Juntang Zhuang, Junlin Yang, Vladimir Iglovikov, Anton Dobrenkii, Madhu Reddiboina, Anubhav Reddy, Xingtong Liu, Cong Gao, Mathias Unberath, Myeonghyeon Kim, Chanho Kim, Chaewon Kim, Hyejin Kim, Gyeongmin Lee, Ihsan Ullah, Miguel Luna, Sang Hyun Park, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. 2018 robotic scene segmentation challenge. *arXiv*, 2020.

Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, and Pablo Arbeláez. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, 106:103726, 2025.

Vivek Singh Bawa, Gurkirt Singh, Francis Kaping'a, Inna Skarga-Bandurova, Elettra Oleari, Alice Leporini, Carmela Landolfo, Pengfei Zhao, Xi Xiang, Gongning Luo, Kuanquan Wang, Liangzhi Li, Bowen Wang, Shang Zhao, Li Li, Armando Stabile, Francesco Setti, Riccardo Muradore, and Fabio Cuzzolin. The SARAS endoscopic surgeon action detection (ESAD) dataset: Challenges and methods. *arXiv*, 2021.

Wenxuan Cheng, Ming Dai, Huimin Lu, and Wankou Yang. Dervos: Decoupling consistent trajectory generation and multimodal understanding for referring video object segmentation. In *CVPR*, 2026.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAMWISE: infusing wisdom in SAM2 for text-driven video segmentation. In *CVPR*, 2025.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023.

Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. Mevis: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11400–11416, 2025.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018.

Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, 2024.

Tim J. M. Jaspers, Ronald L. P. D. de Jong, Yiping Li, Carolus H. J. Kusters, Franciscus H. A. Bakker, Romy C. van Jaarsveld, Gino M. Kuiper, Richard van Hillegersberg, Jelle P. Ruurda, Willem M. Brinkman, Josien P. W. Pluim, Peter H. N. de With, Marcel Breeuwer, Yasmina Al Khalil, and Fons van der Sommen. Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis*, 108:103873, 2026.

Haichao Jiang, Tianming Liang, Wei-Shi Zheng, and Jian-Fang Hu. Refer-agent: A collaborative multi-agent system with reasoning and reflection for referring video object segmentation. In *CVPR*, 2026.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018.

Joël L. Lavanchy, Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Paolo Fiorini, Beat P. Müller-Stich, Philipp C. Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19(11):2249–2257, 2024.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. Referdino: Referring video object segmentation with visual grounding foundations. In *ICCV*, 2025.

Tianming Liang, Haichao Jiang, Yuting Yang, Chaolei Tan, Shuai Li, Wei-Shi Zheng, and Jian-Fang Hu. Long-rvos: A comprehensive benchmark for long-term referring video object segmentation. In *CVPR*, 2026.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: generalized referring expression segmentation. In *CVPR*, 2023.

Haofeng Liu, Mingqi Gao, Xuxiao Luo, Ziyue Wang, Guanyi Qin, Junde Wu, and Yueming Jin. Resurgsam2: Referring segment anything in surgical video via credible long-term tracking. In *MICCAI*, 2025.

Haofeng Liu, Ziyue Wang, Alex Y. W. Kong, Guanyi Qin, Yunqiu Xu, Chang Han Low, Mingqi Gao, Lap Yan Lennon Chan, and Yueming Jin. Unisurgsam: A unified promptable model for reliable surgical video segmentation. *arXiv*, 2026.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. SOC: semantic-assisted object cluster for referring video object segmentation. In *NeurIPS*, 2023.

Lena Maier-Hein, Martin Wagner, Tobias Roß, Annika Reinke, Sebastian Bodenstedt, Peter M. Full, Hellena Hempe, Diana Mîndroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Anna Kisilenko, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Minu Tizabi, Matthias Eisenmann, Tim J. Adler, Janek Gröhl, Melanie Schellenberg, Silvia Seidlitz, T. Y. Emmy Lai, Veith Roethlingshoefer, Fabian Both, Sebastian Bittel, Marc Mengler, Martin Apitz, Stefanie Speidel, Hannes Götz Kenngott, and Beat P. Müller-Stich. Heidelberg colorectal data set for surgical data science in the sensor operating room. *Scientific Data*, 8(1):101, 2021.

Aditya Murali, Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. *arXiv*, 2023.

Feiyu Pan, Hao Fang, Fangkai Li, Yanyu Xu, Yawei Li, Luca Benini, and Xiankai Lu. Semantic and sequential alignment for referring video object segmentation. In *CVPR*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

Dimitrios Psychogyios, Emanuele Colleoni, Beatrice van Amsterdam, Chih-Yang Li, Shu-Yu Huang, Yuchong Li, Fucang Jia, Baosheng Zou, Guotai Wang, Yang Liu, Maxence Boels, Jiayu Huo, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, Sébastien Ourselin, Mengya Xu, An Wang, Yanan Wu, Long Bai, Hongliang Ren, Atsushi Yamada, Yuriko Harai, Yuto Ishikawa, Kazuyuki Hayashi, Jente Simoens, Pieter DeBacker, Francesco Cisternino, Gabriele Furnari, Alexandre Mottrie, Federica Ferraguti, Satoshi Kondo, Satoshi Kasai, Kousuke Hirasawa, Soohee Kim, Seung Hyun Lee, Kyu Eun Lee, Hyoun-Joong Kong, Kui Fu, Chao Li, Shan An, Stefanie Krell, Sebastian Bodenstedt, Nicolás Ayobi, Alejandra Pérez, Santiago Rodríguez, Juanita Puentes, Pablo Arbeláez, Omid Mohareri, and Danail Stoyanov. SAR-RARP50: segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

Tobias Rueckert, Raphaela Maerkl, David Rauber, Leonard Klausmann, Max Gutbrod, Daniel Rueckert, Hubertus Feussner, Dirk Wilhelm, and Christoph Palm. Video dataset for surgical phase, keypoint, and instrument recognition in laparoscopic surgery (phakir). *Data in Brief*, 68: 113147, 2026.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, 2023.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020.

Andru Putra Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Martin Wagner, Beat P. Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M. Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Annika Reinke, Carissa Reid, Tong Yu, Armine Vardazaryan, Chinedu Innocent Nwoye, Nicolas Padoy, Xinyang Liu, Eungjoo Lee, Constantin Disch, Hans Meine, Tong Xia, Fucang Jia, Satoshi Kondo, Wolfgang Reiter, Yueming Jin, Yonghao Long, Meirui Jiang, Qi Dou, Pheng-Ann Heng, Isabell Twick, Kadir Kirtaç, Enes Hosgor, Jon Lindström Bolmgren, Michael Stenzel, Björn von Siemens, Long Zhao, Zhenxiao Ge, Haiming Sun, Di Xie, Mengqi Guo, Daochang Liu, Hannes Götz Kenngott, Felix Nickel, Moritz von Frankenberg, Franziska Mathis-Ullrich, Annette Kopp-Schneider, Lena Maier-Hein, Stefanie Speidel, and Sebastian Bodenstedt. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Medical Image Analysis*, 86:102770, 2023.

Hongqiu Wang, Guang Yang, Shichen Zhang, Jing Qin, Yike Guo, Bo Xu, Yueming Jin, and Lei Zhu. Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. *IEEE Transactions on Medical Imaging*, 43(12):4457–4469, 2024.

Zanyi Wang, Dengyang Jiang, Liuzhuozheng Li, Sizhe Dang, Chengzu Li, Harry Yang, Guang Dai, Mengmeng Wang, and Jingdong Wang. Deforming videos to masks: Flow matching for referring video segmentation. In *ICLR*, 2026.

Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *MICCAI*, 2022.

Meng Wei, Kun Yuan, Shi Li, Yue Zhou, Long Bai, Nassir Navab, Hongliang Ren, Hong Joo Lee, Tom Vercauteren, and Nicolas Padoy. Where it moves, it matters: Referring surgical instrument segmentation via motion. In *AAAI*, 2026.

Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *ICCV*, 2023.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022.

Jihun Yoon, Jiwon Lee, Sunghwan Heo, Hayeong Yu, Jayeon Lim, Chihyun Song, SeulGi Hong, Seungbum Hong, Bokyung Park, SungHyun Park, Woo Jin Hyung, and Min-Kook Choi. hsdb-instrument: Instrument localization database for laparoscopic and robotic surgeries. In *MICCAI*, 2021.

Linfeng Yuan, Miaojing Shi, Zijie Yue, and Qijun Chen. Losh: Long-short text joint prediction network for referring video object segmentation. In *CVPR*, 2024.

Zijian Zhou, Oluwatosin Alabi, Meng Wei, Tom Vercauteren, and Miaojing Shi. Text promptable surgical instrument segmentation with vision-language models. In *NeurIPS*, 2023.

Aneeq Zia, Max Berniker, Rogerio Nespolo, Conor Perreault, Ziheng Wang, Benjamin Mueller, Ryan Schmidt, Kiran Bhattacharyya, Xi Liu, and Anthony M. Jarc. Surgical visual understanding (surgvu) dataset. *arXiv*, 2025.

APPENDIX

To better understand our LD-RSVIS and Cascade-RSVIS in this work, we provide additional details, analysis, results, and discussions as follows:

- **A** ***Details of Surgical Instrument Categories and Procedures***
  We provide the details of surgical instrument classes and procedures in LD-RSVIS.
- **B** ***Construction Pipeline of LD-RSVIS and Additional Annotation Examples***
  In this section, we present the detailed construction pipeline of LD-RSVIS and demonstrate more annotation examples.
- **C** ***More Statistics and Annotation Quality Analysis***
  We show additional statistics and annotation quality analysis of LD-RSVIS in this section.
- **D** ***Details of Loss Function for Optimization***
  We provide details of the loss function for optimization of Cascade-RSVIS.
- **E** ***Model Complexity of Cascade-RSVIS***
  We provide complexity analysis of Cascade-RSVIS in this section.
- **F** ***Additional Results and Analysis***
  We provide additional results and analysis in this section.
- **G** ***Limitation of Cascade-RSVIS***
  This section discusses the limitation of Cascade-RSVIS.
- **H** ***Dataset Specification***
  We discuss dataset specifications of LD-RSVIS.

A DETAILS OF SURGICAL INSTRUMENT CATEGORIES AND PROCEDURES

LD-RSVIS aims at facilitating the development of general RSVIS. To this end, LD-RSVIS provides diverse surgical instrument classes and covers rich surgical procedures as described below.

A.1 SURGICAL INSTRUMENT CATEGORIES

Specifically, LD-RSVIS contains 30 instrument classes, including "*Bipolar Forceps*", "*Prograsp Forceps*", "*Large Needle Driver*", "*Monopolar Curved Scissors*", "*Vessel Sealer*", "*Grasping Retractor*", "*Ultrasound Probe*", "*Suction Instrument*", "*Clip Applier*", "*Grasper*", "*Electric Hook*", "*Irrigator*", "*Scissors*", "*Snare*", "*Palpation Probe*", "*Trocar*", "*HF Coagulation Probe*", "*Needle Probe*", "*Drainage Instrument*", "*Argon Beamer*", "*Hook Clamp*", "*Overholt Clamp*", "*Sponge Clamp*", "*LigaSure*", "*Dissecting Forceps*", "*Grasping Forceps*", "*Suturing Needle*", "*Clip*", "*Catheter*", and "*Needle Holder*".

A.2 SURGICAL PROCEDURES

LD-RSVIS covers 25 procedure categories, including "*Appendectomy*", "*Cardiomyotomy*", "*Cholecystectomy*", "*Colectomy*", "*Colon Cancer Resection*", "*Colorectal Resection*", "*Esophagectomy*", "*Fundoplication*", "*Gastrectomy*", "*Gastrojejunostomy*", "*Heller Myotomy*", "*Hemicolectomy*", "*Hernia Repair*", "*Ladd's Procedure*", "*Liver Resection*", "*Rectal Resection*", "*Rectal Prolapse Repair*", "*Rectopexy*", "*Sigmoidectomy*", "*Splenectomy*", "*Hysterectomy*", "*Nephrectomy*", "*Roux-en-Y Gastric Bypass*", "*Radical Prostatectomy*", and "*Proctocolectomy*".

B CONSTRUCTION PIPELINE OF LD-RSVIS AND ADDITIONAL ANNOTATION EXAMPLES

The construction of LD-RSVIS contains five steps: **(i)** ***Overall design.*** We first determine the scope of referring surgical video instrument segmentation and the criteria for selecting source datasets, procedures, instrument classes, and referring settings. **(ii)** ***Video acquisition.*** We then collect videos from 15 public source datasets, covering surgical scene understanding, phase and action recognition, workflow analysis, general surgical visual understanding, and surgical video pre-training. **(iii)** ***Video***

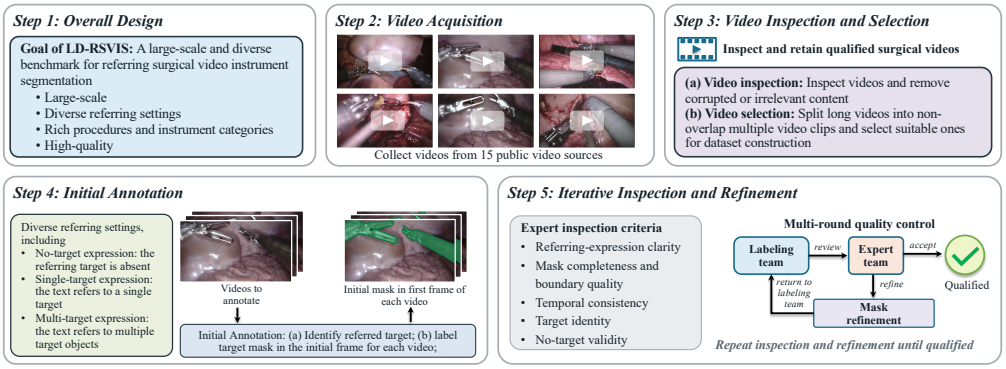


Figure 5: Illustration of the construction pipeline of LD-RSVIS.

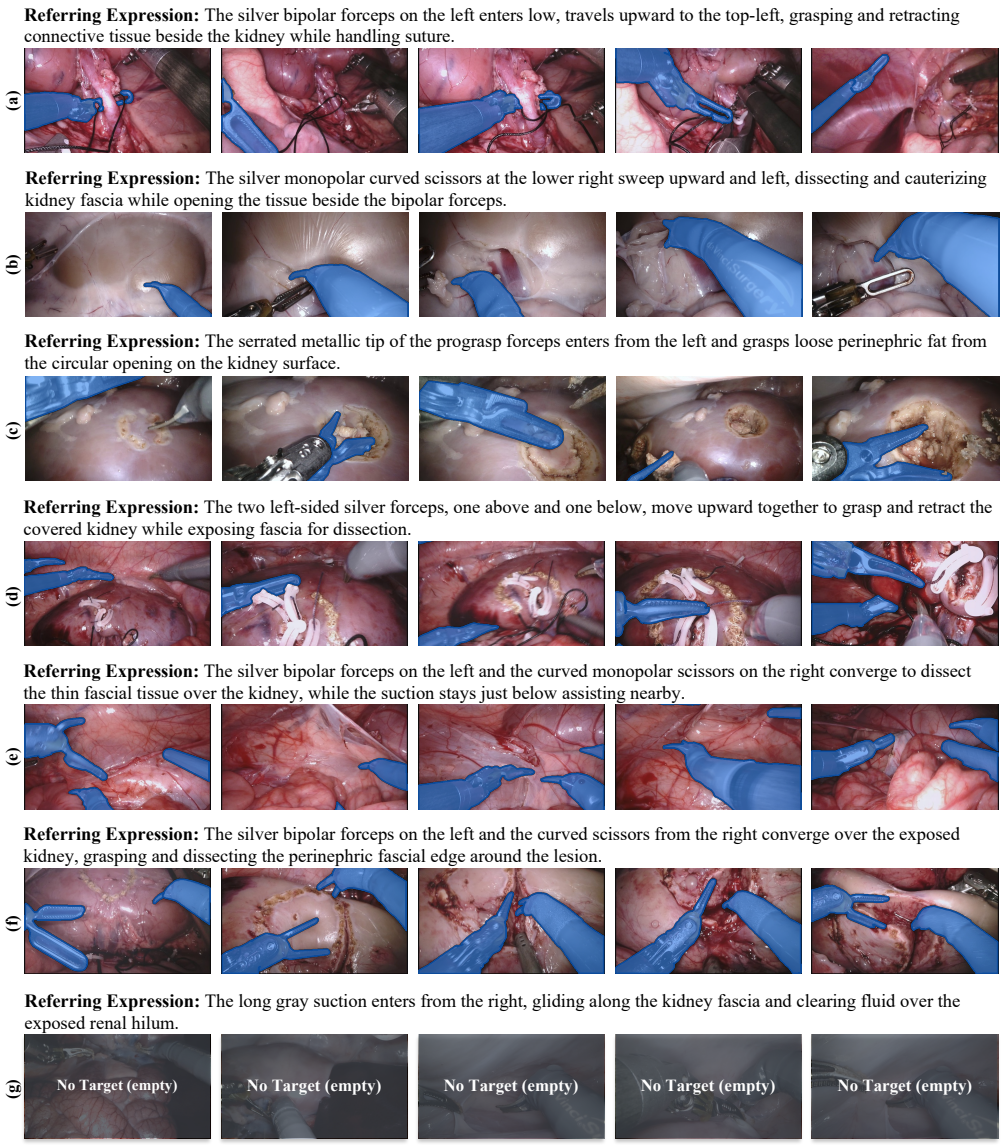


Figure 6: Additional annotation examples in LD-RSVIS.

Table 9: Composition of selected data from source datasets for constructing LD-RSVIS.

Source Dataset	Task of the Source Dataset	Selected Videos	Selected Frames
EndoVis17 (Allan et al., 2019)	surgical scene segmentation and/or detection	18	3,000
EndoVis18 (Allan et al., 2020)	surgical scene segmentation and/or detection	15	2,235
Cholec80 (Twinanda et al., 2017)	surgical workflow understanding	318	98,620
Endoscapes2023 (Murali et al., 2023)	surgical scene segmentation and/or detection	181	55,840
HeiChole (Wagner et al., 2023)	surgical workflow understanding	170	52,460
PhaKIR (Rueckert et al., 2026)	surgical phase and action recognition	60	18,640
AutoLaparo (Wang et al., 2022)	surgical scene segmentation and/or detection	218	67,420
MultiBypass140 (Lavanchy et al., 2024)	surgical phase and action recognition	654	202,350
HeiCo (Maier-Hein et al., 2021)	surgical scene segmentation and/or detection	442	136,840
ESAD (Bawa et al., 2021)	surgical phase and action recognition	169	52,180
GraSP (Ayobi et al., 2025)	surgical scene segmentation and/or detection	238	73,420
SAR-RARP50 (Psychogyios et al., 2024)	surgical phase and action recognition	49	14,210
SurgVU (Zia et al., 2025)	surgical task and activity recognition	393	121,360
hSDB-instrument (Yoon et al., 2021)	surgical scene segmentation and/or detection	115	35,480
SurgeYoutube (Jaspers et al., 2026)	large-scale surgical video pre-training	496	153,307
Total	–	**3,536**	**1,09 Million**

***selection*.** We carefully inspect videos from different sources and remove corrupted or irrelevant content. Qualified long videos are divided into representative continuous sequences that preserve diverse instrument appearances, surgical stages, and temporal conditions while avoiding excessive temporal redundancy. **(iv)** ***Initial annotation*.** For each retained sequence, experts generate referring expressions and offer initial frame-level masks. **(v)** ***Iterative inspection and refinement*.** The initial masks are extended to consistent spatio-temporal masklets by the labeling team and then reviewed by experts for mask completeness, boundary quality, temporal consistency, and target identity. Annotations that do not satisfy the quality requirements are returned to the labeling team for refinement. This process is repeated until all annotations are qualified. The construction pipeline is illustrated in Fig. 5. In Fig. 6, we demonstrate more annotation samples of LD-RSVIS.

C More Statistics and Annotation Quality Analysis

Selected data from source datasets. Our LD-RSVIS comprises 3,536 surgical videos and 1.09 million frames selected from 15 publicly available source datasets. Tab. 9 summarizes the selected data from each source dataset and their contributions to LD-RSVIS. The source pool includes both standard surgical video benchmarks and datasets developed for broader surgical video understanding, providing diverse coverage of surgical procedures, instrument categories, and temporal contexts.

Distribution of different types of expressions. Tab. 10 summarizs the expression composition of LD-RSVIS. Specifically, the training set LD-RSVIS$_{\text{Tra}}$ contains 11,412 multi-target expressions (MTE), 28,171 single-target expressions (STE), and 7,917 no-target expressions (NTE). The test set LD-RSVIS$_{\text{Tst}}$ contains 2,732 MTE, 6,573 STE, and 1,861 NTE. Together, our LD-RSVIS comprises 58,666 referring expressions, with 34,744, 14,144, and 9,778 STE, MTE, and NTE, respectively.

Table 10: Distribution of different types of referring expression in LD-RSVIS.

	MTE	STE	NTE	Total
LD-RSVIS$_{\text{Tra}}$	11,412	28,171	7,917	47,500
LD-RSVIS$_{\text{Tst}}$	2,732	6,573	1,861	11,166
LD-RSVIS	14,144	34,744	9,778	58,666

Annotation quality analysis. To analyze our annotation quality, we randomly choose 50 sequences from our LD-RSVIS and ask an independent group of external experts to inspect and re-label them. Then, we compute the Intersection over Union (IoU) of new and original masklets. The IoU of these selected sequences is 0.93, validating the quality and reliability of our annotations.

D Details of Loss Function for Optimization

Cascade-RSVIS contains two stages of TCM and CGS. Given a video sequence $\mathcal{I} = \{I_i\}_{i=1}^{L}$ with L frames and a referring expression, both stages predict masks. During training, given the groundtruth, we follow SAM 2 (Ravi et al., 2025) and utilize multiple losses, including focal loss, dice loss, IoU loss, and binary cross entropy loss to compute the frame-level loss $\mathcal{L}^{i}_{\text{TCM}}$ and $\mathcal{L}^{i}_{\text{CGS}}$ for TCM and CGS, respectively (please refer to these losses in (Ravi et al., 2025) for details). The losses over the

Table 11: Comparison on existing public RSVIS benchmarks. oIoU and mIoU denote Overall IoU and Mean IoU, respectively.

	EndoVis-RS						Ref-EndoVis (tool)					
	RS17			RS18			Ref17			Ref18		
Methods	oIoU	mIoU	mAP	oIoU	mIoU	mAP	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
VIS-Net (Wang et al., 2024)[T-MI'24]	65.5	70.1	53.8	74.2	72.3	60.1	61.4	61.1	61.2	68.6	68.2	68.4
SurgRef (Wei et al., 2026)[AAAI'26]	66.4	70.8	54.7	74.8	73.4	60.8	70.8	69.6	70.2	76.1	74.7	75.4
ReSurgSAM2 (Liu et al., 2025)[MICCAI'25]	68.7	72.9	57.2	76.5	75.3	62.9	77.8	77.7	77.7	80.9	80.3	80.6
ReferMo (Liang et al., 2026)[CVPR'26]	69.4	73.4	57.6	77.2	75.8	63.3	78.2	78.8	78.5	81.0	81.4	81.2
FlowRVS (Wang et al., 2026)[ICLR'26]	69.0	74.0	58.1	76.8	76.4	64.0	77.8	80.6	79.2	80.8	82.8	81.8
Cascade-RSVIS (ours)	**71.6**	**75.9**	**59.8**	**79.0**	**78.1**	**65.3**	**80.2**	**81.0**	**80.6**	**82.1**	**83.1**	**82.6**

entire video sequence are computed as $\mathcal{L}_{\text{TCM}} = \sum_{i=1}^{L} \mathcal{L}_{\text{TCM}}^{i}$ and $\mathcal{L}_{\text{CGS}} = \sum_{i=1}^{L} \mathcal{L}_{\text{CGS}}^{i}$. In this work, we adopt a stage-wise optimization strategy. Specifically, we first optimize TCM using $\mathcal{L}_{\text{TCM}}$. After this, CGS is optimized using $\mathcal{L}_{\text{CGS}}$ with the same input together with outputs from TCM.

E Model Complexity of Cascade-RSVIS

We train Cascade-RSVIS using two NVIDIA A100 (40G) GPUs for around 20 hours. The resolution of video frames for both training and inference is 896×896. Cascade-RSVIS contains 292M total parameters. During inference, the video is divided into eight-frame clips and processed on a single NVIDIA A100 GPU. The peak GPU memory requirement for running Cascade-RSVIS is 16 GB, and its inference speed is 25 frames per second (*fps*).

F Additional Results and Analysis

F.1 Comparison on Existing RSVIS Benchmarks

To evaluate the generalization ability of Cascade-RSVIS on existing public surgical RSVIS benchmarks, we conduct experiments on EndoVis-RS (Wang et al., 2024) and Ref-EndoVis (Liu et al., 2025), which are the two publicly available benchmarks with released data. We compare Cascade-RSVIS with three surgical RSVIS methods, including VIS-Net (Wang et al., 2024), SurgRef (Wei et al., 2026), and ReSurgSAM2 (Liu et al., 2025), as well as two recent general RVOS methods, ReferMo (Liang et al., 2026) and FlowRVS (Wang et al., 2026). For Ref-EndoVis, we report results on tool expressions only, since LD-RSVIS focuses on surgical instruments and their parts rather than tissues. We follow the original evaluation protocols of each benchmark, adopting Overall IoU (oIoU), Mean IoU (mIoU), and mAP for EndoVis-RS, and $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$ for Ref-EndoVis.

Tab. 11 reports the comparison results. ReferMo and FlowRVS transfer favorably to both surgical benchmarks and outperform the surgical RSVIS methods in terms of mAP and $\mathcal{J}\&\mathcal{F}$ across all four subsets, with FlowRVS achieving the strongest aggregate results among the compared approaches. Their margins over ReSurgSAM2 are comparatively smaller on these benchmarks than under the STE setting of LD-RSVIS. This is because the existing benchmarks focus on whole-instrument targets in a limited number of surgical scenes, which closely matches the target-initialization and long-term tracking design of ReSurgSAM2, whereas LD-RSVIS additionally includes fine-grained part-level expressions and substantially greater scene diversity. Nevertheless, Cascade-RSVIS obtains mAP scores of 59.8% and 65.3% on RS17 and RS18 of EndoVis-RS, exceeding FlowRVS by 1.7% and 1.3%, respectively. On Ref-17 and Ref-18 of Ref-EndoVis (tool), it achieves $\mathcal{J}\&\mathcal{F}$ scores of 80.6% and 82.6%, improving on FlowRVS by 1.4% and 0.8%. These results consistently demonstrate the effectiveness of Cascade-RSVIS for referring surgical video instrument segmentation.

F.2 Qualitative Comparison

To provide further qualitative analysis of Cascade-RSVIS, we compare its segmentation results with four representative RSVIS methods in Fig. 7. Fig. 7-left shows a challenging single-target example, where the referred instrument tip occupies only a small region and moves substantially across the video. VIS-Net tends to segment the complete parent instrument and switches to a neighboring instrument in the late frame, ReSurgSAM2 exhibits leakage into the adjacent instrument part, and

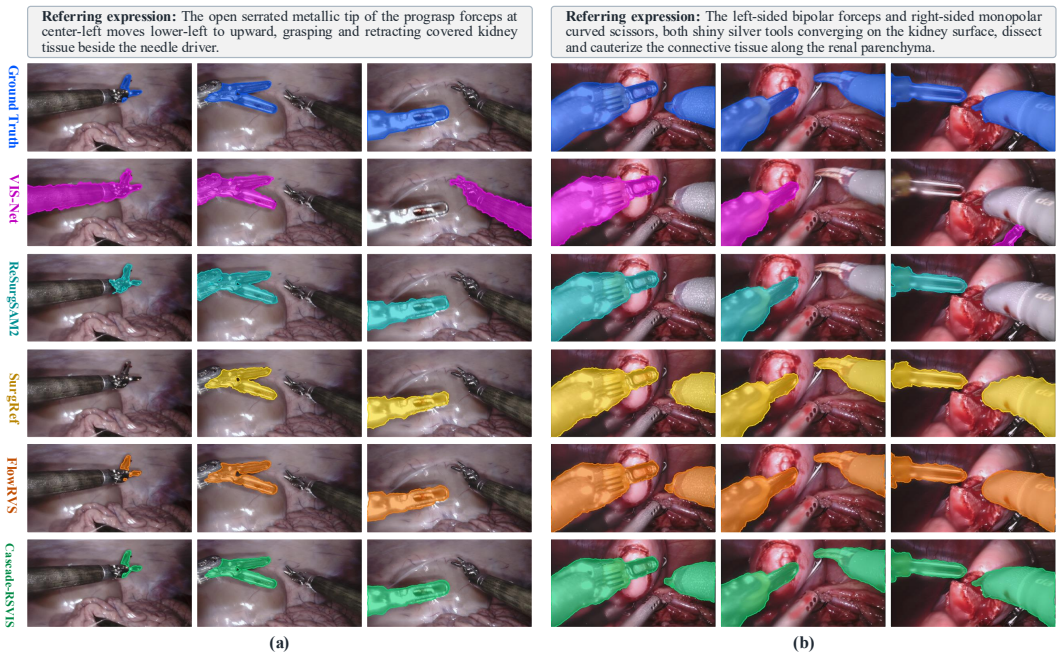


Figure 7: Qualitative comparison of Cascade-RSVIS with representative methods on LD-RSVIS. Image (a) shows the results of a single target case, while image (b) shows a multi-target case.

SurgRef misses the target in the early frame. FlowRVS localizes the referred part more accurately but still exhibits local boundary deviations. Figure 7-right further shows a multi-target example in which two instruments are referred to among four simultaneously visible instruments. The compared methods either omit one referred target, switch to an unreferred distractor, or lose fine instrument regions across frames; FlowRVS retains both target identities but loses part of the fine regions in the late frame. In contrast, Cascade-RSVIS consistently identifies the referred target set and produces accurate masks throughout both videos, demonstrating the effectiveness of target-context mining and context-guided spatial-temporal segmentation.

G Limitation of Cascade-RSVIS

In this work, we propose a simple yet effective Cascade-RSVIS to encourage more future research on RSVIS. Despite promising results, Cascade-RSVIS has two limitations. First, it may degrade in highly challenging surgical scenes. In these scenes, very small instrument parts may be heavily occluded, and multiple instruments with similar appearances may occur simultaneously, resulting in lower segmentation performance. To handle these cases, special strategies such as leveraging richer spatial context through explicit instrument-instrument interaction or target trajectory cues are needed. Second, the current framework does not explicitly model instrument motion. Although temporal information is implicitly captured through video features, explicit motion cues, such as motion-aware correspondence or instrument trajectories, may provide complementary information for distinguishing visually similar instruments and maintaining target consistency under rapid movement or occlusion. Exploring such motion-aware modeling is an interesting direction for future work. Considering that our current goal is to provide a feasible baseline for RSVIS, we leave the above explorations to future work.

H Dataset Specification

Annotator Protection. To safeguard annotator well-being, all participants provide informed consent and are notified in advance that the surgical videos may contain potentially distressing content. Annotators are allowed to skip any video that causes discomfort and may withdraw from the annotation process at any time.

Maintenance. LD-RSVIS will be hosted on GitHub and/or Hugging Face. This allows us to conveniently check feedback from the community and improve LD-RSVIS through necessary maintenance and updates. In addition, all the annotations and results in this work will be made publicly available. Our ultimate goal is to offer a long-term and stable platform for the community to foster research on referring surgical video instrument segmentation.

Responsible Use of LD-RSVIS. LD-RSVIS is intended to support research on referring surgical video instrument segmentation and is released for research purposes only. Users should be aware that, due to potential biases inherited from the source datasets and the annotation process, the dataset may exhibit imbalances across surgical procedures, institutions, acquisition devices, and instrument categories. It is worth noting that, the clinical validation of this work is required before their deployment in real-world surgical settings.

Potential Limitation. A potential limitation of LD-RSVIS is the substantial cost and complexity of annotation. Accurately segmenting referred surgical instruments across video sequences, in diverse multi-target, single-target and no-target settings, requires detailed spatiotemporal annotations. This process is labor-intensive and time-consuming, making it difficult to scale the benchmark rapidly to more videos, surgical procedures, and clinical domains.