DISCOVER AND VERIFY: FINE-GRAINED VIDEO OBJECT SEGMENTATION WITH WEAK-TEXT PROMPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-guided video segmentation converts text queries into spatial prompts, reducing the need for manual spatial annotations. However, this paradigm commonly relies on an implicit assumption that semantic similarity is sufficient to determine the identity of the target. This assumption often fails in fine-grained settings where only an easily available category name is provided as a weak-text prompt. Visually similar objects from different categories may receive similar semantic similarity scores, making it difficult to reliably distinguish the target from confusing objects. To address this issue, we propose a selective weak-text prompted fine-grained segmentation framework based on a "discover-and-verify" mechanism. Confuser-aware Prompt Discriminative Grounding (CPDG) compares the target category with potential confuser categories to learn latent discriminative cues. These cues are then used to locate and aggregate local discriminative evidence within candidate regions, allowing CPDG to rerank the candidate regions. Semantic Consistency Verification (SCV) further combines semantic and visual evidence to verify the top-ranked candidate and rejects the candidate hypothesis when the available evidence is insufficient. Extensive experiments on two weak-text fine-grained datasets, EndoVis17 and EndoVis18, demonstrate the effectiveness of our method, which consistently improves weak-text prompted fine-grained segmentation performance and improves mcIoU, our primary evaluation metric, over the strongest compared method by 1.31% and 16.91%, respectively.

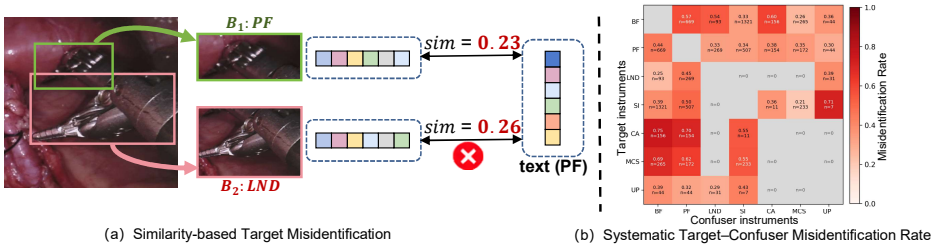
1 INTRODUCTION

Recent advances in vision-language foundation models (Radford et al., 2021a; Wang et al., 2022; Lüddecke & Ecker, 2022) have made it possible to localize and segment objects specified by natural language. Language grounding models such as GroundingDINO (Liu et al., 2024) can localize objects from text queries, while promptable video segmentation models (Ravi et al., 2025) can segment and propagate an object through a video from sparse visual prompts. These capabilities support a grounding-based promptable segmentation pipeline: a text query is first grounded into an object hypothesis, which is then converted into a spatial prompt for mask generation (Ren et al., 2024) and temporal propagation (Ravi et al., 2025). By substituting manually defined spatial prompts with text-based grounding, this pipeline lowers user spatial interaction demands, suiting medical and surgical videos where annotation relies on expert knowledge and heavy manual work.

Several studies have explored this grounding-based promptable segmentation pipeline. The Grounded-SAM family connects language grounding models with promptable segmentation models to generate object masks from text queries. More recent methods, such as SAMWISE (Cuttano et al., 2025) and ReferDINO (Liang et al., 2025), further combine language representations with temporal modeling to support text-guided video object segmentation. Most existing methods identify targets through semantic similarity established by cross-modal correspondence between the text query and candidate visual regions, and select the best-matched one. They generally assume relatively rich descriptions containing appearance, motion, or relational cues that help resolve ambiguity (Yuan et al., 2024). These methods therefore implicitly assume that the resulting matching scores provide sufficient separation to rank the target region above other candidate regions.

However, requiring users to provide a detailed description for every target increases interaction effort, particularly in specialized domains where such descriptions may require expert knowledge.

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.



(a) Similarity-based Target Misidentification (b) Systematic Target–Confuser Misidentification Rate

Figure 1: Motivation: semantic similarity alone is insufficient for fine-grained target discrimination. (a) A representative failure case where the confuser (Large Needle Driver, LND) receives a higher semantic similarity score than the true target (Prograsp Forceps, PF). (b) Dataset-level statistics on EndoVis18 show that this phenomenon occurs systematically across multiple instrument pairs. These results indicate that semantic similarity alone may remain unable to distinguish visually similar categories in fine-grained video segmentation with weak-text prompts.

Category names provide a simpler alternative because users do not need to manually describe the target's appearance, motion, or relations. We refer to a category-name-only query as a *weak-text prompt*. Consequently, *the target and a visually similar non-target object, which we refer to as a confuser, may receive similar text-region semantic similarity scores*. Figure 1(a) shows a representative failure case in which a confuser (LND) receives a higher semantic similarity score than the true target (PF). Figure 1(b) further shows that this phenomenon occurs systematically across multiple instrument pairs on EndoVis18: in 51.8% of the evaluated target-confuser pairs, GroundingDINO assigns a higher semantic similarity score to the confuser than to the true target. For the evaluated model and dataset, this result shows that semantic similarity may remain unable to distinguish visually similar categories. This observation raises a central question: **what local category-discriminative cues within a candidate region can support the target while separating it from plausible confusers?**

We approach fine-grained video segmentation from weak-text prompts as a process of discriminative evidence discovery and hypothesis verification. Instead of asking only which candidate is most semantically similar to the text query, we consider two questions: (1) which local category-discriminative cues distinguish the target from its confusers, and (2) whether the discovered discriminative cues are reliable enough to support the resulting target hypothesis. The first question requires the model to identify visual local category-discriminative cues that increase the separation between the target and confuser categories. The second requires the model to assess whether the selected candidate contains sufficient supporting evidence. When such evidence is insufficient, the model should reject the candidate rather than pass an unreliable spatial prompt to the downstream segmentation and propagation model. Based on this idea, we propose a weak-text-prompted fine-grained video object segmentation framework with a discover-and-verify (DaV) strategy. The Confuser-aware Prompt Discriminative Grounding module (CPDG) compares the target with confuser categories and learns latent attribute representations that capture local category-discriminative cues without manual part annotations. These cues are aggregated to score and rerank candidate objects according to their support for the target category. The Semantic Consistency Verification module (SCV) then treats the selected candidate as a hypothesis and verifies it using semantic and visual evidence, rejecting candidates with insufficient support. Thus, CPDG identifies the most discriminative candidate, while SCV determines whether it is reliable enough to accept.

We evaluate DaV on the EndoVis17 and EndoVis18 surgical video benchmarks, where visually similar instrument categories provide a challenging testbed for weak-text prompt fine-grained segmentation. Our experiments demonstrate consistent improvements in fine-grained target discrimination, segmentation, and absent-query rejection. Our contributions are summarized as follows:

- We study weak-text fine-grained video object segmentation and show that semantic similarity alone is insufficient for target-confuser separation.
- We propose a discover-and-verify framework, where CPDG models target-confuser differences to discover local discriminative cues and rerank candidate hypotheses.
- We introduce SCV to verify candidate hypotheses and reject those with insufficient semantic and visual support.

2 Related Work

Language-Guided and Promptable Segmentation. Vision-language pretraining enables models to align text with visual regions for language-guided localization and segmentation. CLIP (Radford et al., 2021a; Yue et al., 2023) learns transferable cross-modal representations, while GroundingDINO (Liu et al., 2024) extends this ability to text-conditioned spatial localization. Promptable segmentation models such as SAM (Kirillov et al., 2023) and SAM 2 (Ravi et al., 2025) further enable mask prediction from spatial prompts. Their combination forms a common text-guided segmentation paradigm in which language queries are first grounded into spatial prompts and then segmented. However, this paradigm often assumes sufficiently informative text; category-name-only prompts may provide inadequate cues for distinguishing visually similar fine-grained categories.

Text-promptable segmentation. Text-promptable segmentation localizes image regions from natural-language queries and is closely related to referring expression segmentation. Earlier methods use CNN/RNN-based multimodal fusion, while later approaches adopt attention and transformers for stronger cross-modal interaction (Yu et al., 2018; Ye et al., 2019; Ding et al., 2021; Wang et al., 2022; Lüddecke & Ecker, 2022). Pretrained vision-language models further extend this paradigm toward open-vocabulary and interactive segmentation (Radford et al., 2021b; Wang et al., 2022; Lüddecke & Ecker, 2022). In surgical imaging, TP-SIS uses class names, prompt templates, and LLM-generated descriptions (Zhou et al., 2023), while subsequent work highlights the reliance of many methods on oracle presence information or richer prompts (Choi & Park, 2024). Referring surgical segmentation has also used richer spatial or motion expressions (Wang et al., 2024). In contrast, we consider category-name-only weak text prompts without handcrafted descriptions or additional contextual cues.

Fine-Grained Recognition and Discriminative Evidence Learning. Fine-grained visual recognition distinguishes categories with subtle visual differences by exploiting localized discriminative regions and category-specific attributes. Part-based methods such as Part-RCNN (Huang et al., 2016) and attention-based approaches (Zheng et al., 2017) learn discriminative object parts, while later studies further improve inter-class separation through attribute learning and contrastive objectives. However, these methods mainly target image-level classification. Weak-text video segmentation additionally requires spatial candidate localization, invalid-hypothesis rejection, and reliable prompt generation. Our work extends discriminative evidence learning into the grounding stage by explicitly modeling target-confuser relationships and verifying candidate consistency.

3 Method

We first define the weak-text-prompted fine-grained segmentation task (Section 3.1). We then describe the Confuser-aware Prompt Discriminative Grounding (CPDG) module (Section 3.2), which mines target-versus-confuser evidence and selects the most strongly supported candidate. Section 3.3 presents Semantic Consistency Verification (SCV), which verifies whether the selected candidate contains sufficient evidence to be accepted or should instead be rejected. Finally, Section 3.4 describes how the verified spatial prompt is passed to a frozen promptable video segmentation model for mask generation and temporal propagation.

3.1 Task Definition

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input frame and T a weak text prompt containing only the name of a target category c (e.g., "Bipolar Forceps"). The prompt is encoded as $q_c = E_{\text{text}}(T) \in \mathbb{R}^d$, where E_{text} is the text encoder and d is the common embedding dimension. A grounding model (e.g., GroundingDINO (Liu et al., 2024)) produces a set of K candidate boxes $\mathcal{B} = \{B_1, \ldots, B_K\}$ and their coarse grounding scores $\{s_i^{\text{coa}}\}_{i=1}^K$. Conventional grounding performs forced selection:

$$i^\star_{\text{base}} = \underset{i\in\{1,\ldots,K\}}{\arg\max}\ \text{sim}(q_c, g_i), \tag{1}$$

where g_i denotes the grounding representation of candidate B_i. This forced-selection strategy can be unreliable in fine-grained settings with category-name-only (weak text) prompts. Different categories may share similar visual structures, while the category name provides limited information about the local cues that distinguish the target from its plausible confusers. Consequently, a non-target candidate may receive a similarity score comparable to or even higher than that of the true

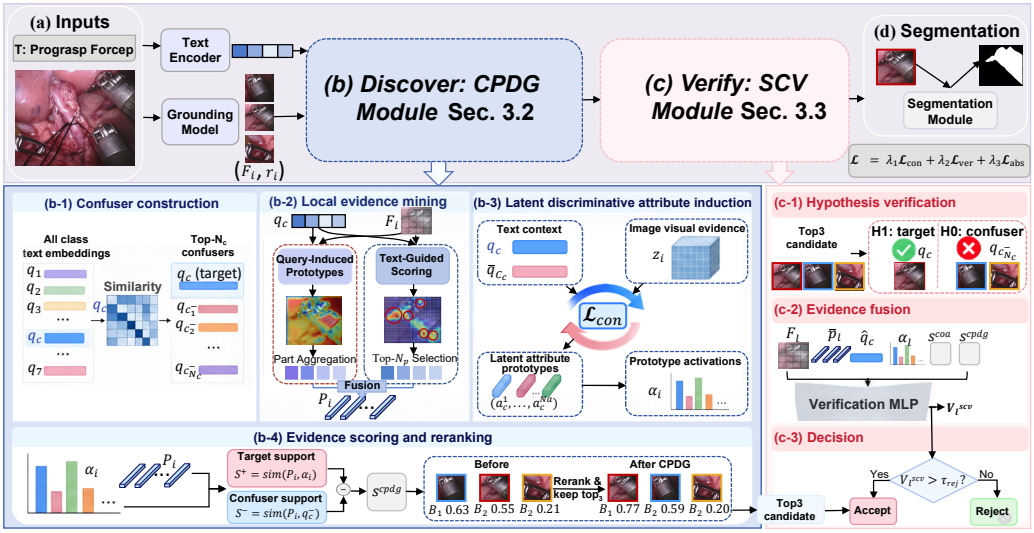


Figure 2: Overall architecture of the proposed DaV framework.

target. More importantly, conventional grounding must still select a candidate when the queried category is absent. We instead formulate the final output as a selective decision. Let V_i denote the verification logit predicted by SCV for candidate B_i. The candidate with the strongest verified support is:

$$i^\star = \underset{i\in\{1,\dots,K\}}{\arg\max}\ V_i. \tag{2}$$

The final prediction is:

$$\hat{y} = f_\theta(I, q_c, \mathcal{B}) = \begin{cases} i^\star, & V_{i^\star} > \tau_{\text{rej}}, \\ \varnothing, & V_{i^\star} \le \tau_{\text{rej}}, \end{cases} \tag{3}$$

where τ_{rej} is the rejection threshold. The output $\hat{y} = i^\star$ indicates that candidate $B_{i^\star}$ is accepted as the target, whereas $\hat{y} = \varnothing$ indicates that the model abstains. This formulation requires the model not only to rank candidate regions but also to determine whether the strongest candidate contains sufficient evidence for the queried category.

3.2 Confuser-aware Prompt Discriminative Grounding

CPDG performs discriminative evidence discovery through five operations: candidate encoding, discriminative evidence localization, confuser construction, discriminative attribute induction, and evidence-aware candidate scoring. Rather than treating each candidate as a target object, CPDG searches for local discriminative cues that support the target category while remaining inconsistent with plausible confusers.

Candidate encoding. Each candidate box B_i is cropped and resized to obtain an ROI R_i. A trainable ROI encoder $E_{\text{roi}}(\cdot)$ produces a global candidate representation $r_i \in \mathbb{R}^d$ and a spatial feature map $F_i \in \mathbb{R}^{d\times h\times w}$ via $(r_i, F_i) = E_{\text{roi}}(R_i)$, where h and w are the spatial dimensions of the ROI feature map.

Discriminative evidence localization. To determine *where* the discriminative evidence lies inside a coarse box, we perform dual-path part mining attention, which includes a query-gated sparse localization (QGSL) branch and a query-induced prototype aggregation (QIPA) branch. QGSL extracts sharply localized responses, whereas QIPA aggregates spatially distributed evidence using query-induced prototypes. Figure 3 illustrates how local discriminative evidence is mined from each candidate using the proposed dual-path part mining attention.

Query-Gated Sparse Localization. We first project the query q_c into the visual channel space to generate a channel-wise gate $\gamma_c = \sigma(W_g q_c) \in \mathbb{R}^d$, which recalibrates the ROI feature map F_i and produces a query-conditioned saliency map A_i:

$$A_i(u,v) = \underset{u,v}{\text{softmax}}\left[w_s^\top \phi(F_i(u,v)\odot\gamma_c + W_q q_c)\right], \tag{4}$$

where $F_i(u, v) \in \mathbb{R}^d$ is the feature at spatial position (u, v), W_g and W_q are learnable projections, $\sigma(\cdot)$ is the sigmoid function, w_s is a learnable scoring vector, $\odot$ denotes element-wise multiplication, and $\phi(\cdot)$ is a nonlinear transformation. We select the N_p positions with the highest attention values and gather their features as sparse part tokens:

$$P_i^{\mathrm{sp}} = \mathrm{TopN}_{N_p}(F_i, A_i) \in \mathbb{R}^{N_p \times d}. \quad (5)$$

Query-Induced Prototype Aggregation. To complement sparse localization, we further transform the query q_c into N_p query-specific part prototypes:

$$Q^{\mathrm{pr}} = \mathrm{reshape}(W_p q_c) \in \mathbb{R}^{N_p \times d}, \quad (6)$$

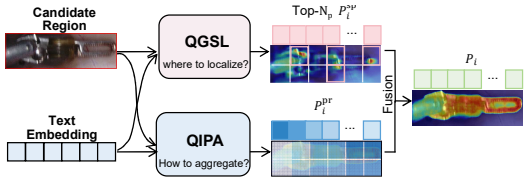


Figure 3: Architecture of the dual-path part mining attention.

and use them to softly aggregate ROI features, where W_p is a learnable projection. Let $\bar{F}_i \in \mathbb{R}^{hw \times d}$ denote the flattened spatial feature map of F_i. The prototype-guided part tokens are computed as:

$$P_i^{\mathrm{pr}} = \mathrm{softmax}\left(\frac{Q^{\mathrm{pr}}(\bar{F}_i W_k)^\top}{\sqrt{d}}\right)(\bar{F}_i W_v) \in \mathbb{R}^{N_p \times d}, \quad (7)$$

where $W_k \in \mathbb{R}^{d \times d}$ and $W_v \in \mathbb{R}^{d \times d}$ are learnable key and value projection matrices, respectively, and $\sqrt{d}$ is the standard scaling factor for stabilizing dot-product attention.

Dual-Path Fusion. Finally, we fuse the sparse localized tokens and the prototype-aggregated tokens to obtain the final query-aware part representation:

$$P_i = \Psi(P_i^{\mathrm{sp}}, P_i^{\mathrm{pr}}) = \{p_i^m\}_{m=1}^{N_p} \in \mathbb{R}^{N_p \times d}, \quad (8)$$

where $\Psi(\cdot)$ denotes a lightweight fusion operator. In this way, the proposed module combines the localization sharpness of sparse attention with the semantic stability of prototype-based aggregation.

Confuser construction. Let $\mathcal{Y}$ denote the category vocabulary. For the queried category c, we construct a set of N_c candidate confusers:

$$\mathcal{C}_c = \{c_1^-, \ldots, c_{N_c}^-\} = \underset{c' \in \mathcal{Y} \setminus \{c\}}{\arg\mathrm{top}_{N_c}} \ \mathrm{sim}(q_c, q_{c'}), \quad (9)$$

where $q_{c'} = E_{\mathrm{text}}(T_{c'})$. These confusers are the categories against which discrimination is informative, and categories far from T_c contribute little evidence.

Latent discriminative attribute induction. The category prompt q_c alone is too coarse to disambiguate T_c from $\mathcal{C}_c$. We induce a small set of *latent attribute prototypes* that capture what visually separates T_c from $\mathcal{C}_c$, without any manual part annotation. Concretely, a lightweight network $\mathcal{D}_{\mathrm{attr}}$ maps the target embedding q_c and the aggregated confuser context into N_a learnable prototypes:

$$\mathcal{A}_c = \mathcal{D}_{\mathrm{attr}}(q_c \,\|\, \bar{q}_{\mathcal{C}_c}) = \{a_c^j\}_{j=1}^{N_a} \in \mathbb{R}^{N_a \times d}, \qquad \bar{q}_{\mathcal{C}_c} = \frac{1}{N_c}\sum_{j=1}^{N_c} q_{c_j^-}, \quad (10)$$

where $\|$ is channel-wise concatenation. The discriminative cue is not derived from q_c alone, but is learned from visual evidence through the target-confuser contrastive objective in Section 3.4.

Candidate attribute activation. The prototypes $\mathcal{A}_c$ are class-level: they are shared by all candidates of category T_c. To determine which attributes are supported by a particular candidate B_i, we first aggregate its part tokens $\bar{p}_i = \frac{1}{N_p}\sum_{m=1}^{N_p} p_i^m$. We then construct a candidate-conditioned evidence representation:

$$z_i = G_{\mathrm{attr}}\big(r_i,\ \bar{p}_i,\ \hat{q}_c\big) \in \mathbb{R}^d, \qquad \hat{q}_c = W_q^{\mathrm{attr}} q_c, \quad (11)$$

where G_{attr} is a lightweight fusion network, $\hat{q}_c$ is the query projects into the visual feature space, and $W_q^{\mathrm{attr}} \in \mathbb{R}^{d \times d}$ is a learnable projection matrix that maps the category embedding q_c into the visual feature space. The logit of the j-th attribute and the candidate attribute vector is

$$\ell_i^{(j)} = \mathrm{sim}\big(z_i,\ a_c^j\big), \qquad \boldsymbol{\ell}_i = \big[\ell_i^{(1)}, \ldots, \ell_i^{(N_a)}\big]^\top. \quad (12)$$

Each entry of $\boldsymbol{\alpha}_i = \sigma(\boldsymbol{\ell}_i) \in [0, 1]^{N_a}$ measures the degree to which the corresponding target-specific attribute is supported by candidate B_i.

Evidence scoring and reranking. For each part token p_i^m, learning from discriminative evidence localization phase, we measure its alignment to the latent discriminative attributes, and penalize its alignment to confuser categories:

$$d_i^m = \max_j \mathrm{sim}(p_i^m, a_c^j) \; - \; \lambda_c \max_{c^- \in \mathcal{C}_c} \mathrm{sim}(p_i^m, q_{c^-}), \tag{13}$$

where λ_c is the confuser-suppression weight. Positive d_i^m therefore denotes *discriminative* evidence: evidence that supports T_c while remaining inconsistent with every confuser. The candidate-level discriminative score aggregates the selected evidence and its confuser counter-evidence:

$$s_i^{\mathrm{cpdg}} = \frac{1}{N_s} \sum_{m \in \mathcal{S}_i} d_i^m + \lambda_{\mathrm{coa}} s_i^{\mathrm{coa}}, \tag{14}$$

where $\mathcal{S}_i$ contains the indices of the N_s part tokens with the highest discriminative scores. CPDG produces a discriminative score, a localized evidence map, and the strongest evidence-aware representation for selecting candidate. These candidate hypotheses are subsequently evaluated by SCV.

3.3 Semantic Consistency Verification

Reranking alone answers which candidate looks most like the target. It does not answer whether enough evidence exists to exclude the confusers. SCV closes this gap by treating B_l as a hypothesis to be verified, not a confidence value to be thresholded.

Candidate hypothesis construction. CPDG retains the indices of the $L = 3$ candidates with the highest discriminative scores:

$$\mathcal{I}_L = \underset{i \in \{1,\dots,K\}}{\arg\mathrm{top}_L} \, s_i^{\mathrm{cpdg}}, \tag{15}$$

where $\mathcal{I}_L \subseteq \{1, \dots, K\}$. The corresponding candidate boxes $\{B_l \mid l \in \mathcal{I}_L\}$ constitute a compact set of target hypotheses for subsequent verification:

$$H_1^{(l)} : \; B_l \text{ belongs to } T_c, \qquad\qquad H_0^{(l)} : \; B_l \text{ belongs to some } c^- \in \mathcal{C}_c. \tag{16}$$

Unlike top-1 verification, this formulation allows SCV to recover a valid lower-ranked candidate when the highest-scoring CPDG proposal is incorrect, while retaining the ability to reject the query when no candidate is sufficiently supported.

Multi-source semantic verification. For each candidate B_l, we construct a verification representation by combining its global appearance r_l, local part representation $\bar{p}_l$, query semantics $\hat{q}_c$, attribute evidence $\boldsymbol{\alpha}_l$, and grounding confidence:

$$u_l = \left[r_l \parallel \bar{p}_l \parallel \hat{q}_c \parallel \boldsymbol{\alpha}_l \parallel s_l^{\mathrm{coa}} \parallel s_l^{\mathrm{cpdg}} \right], l \in \mathcal{I}_L, \tag{17}$$

where the coarse grounding score s_l^{coa} provides the original localization confidence, whereas s_l^{cpdg} summarizes the target-versus-confuser evidence discovered by CPDG. A lightweight verification network maps u_l to a verification logit: $V_l = V_{\mathrm{scv}}(u_l), l \in \mathcal{I}_L$, and the corresponding verification probability is $V_l^{\mathrm{scv}} = \sigma(V_l)$. A larger V_l^{scv} indicates stronger semantic support for $H_1^{(l)}$ over $H_0^{(l)}$.

Verified selection and rejection. SCV selects the candidate with the highest verification probability from the CPDG shortlist: $l^\star = \underset{l \in \mathcal{I}_L}{\arg\max} \, V_l^{\mathrm{scv}}$. The final verified spatial prompt is:

$$B^\star = \begin{cases} B_{l^\star}, & V_{l^\star}^{\mathrm{scv}} > \tau_{\mathrm{rej}}, \\ \varnothing, & V_{l^\star}^{\mathrm{scv}} \le \tau_{\mathrm{rej}}. \end{cases} \tag{18}$$

When $B^\star \neq \varnothing$, the verified box is sent to the downstream segmenter: $M = S(I, B^\star)$. When $B^\star = \varnothing$, the method abstains and does not initialize mask generation or temporal propagation. In this way, CPDG and SCV play complementary roles: CPDG discovers and ranks candidates using target-versus-confuser evidence, whereas SCV corrects the retained ranking and determines whether the strongest candidate should be accepted.

3.4 Training Objective

We jointly optimize CPDG and SCV using target-confuser contrastive, semantic verification, and absent-query rejection objectives. Let B_c^{gt} denote the ground-truth box of category c, and let δ be the IoU threshold used to determine whether a proposal matches the target. For a present-query sample, we define $\mathcal{I}_+ = \{i : \mathrm{IoU}(B_i, B_c^{\mathrm{gt}}) \geq \delta\}, \mathcal{I}_- = \{1, \ldots, K\} \setminus \mathcal{I}_+$.

Target-confuser contrastive loss. To make the induced attribute prototypes discriminative, we require the evidence representation of a positive candidate to align more strongly with the target attributes than with any confuser category:

$$\mathcal{L}_{\mathrm{con}} = \sum_{i \in \mathcal{I}_+} \left[m_{\mathrm{con}} - \max_j \mathrm{sim}(z_i, a_c^j) + \max_{c^- \in \mathcal{C}_c} \mathrm{sim}(z_i, q_{c^-}) \right]_+, \tag{19}$$

where $[x]_+ = \max(0, x)$ and m_{con} is the contrastive margin, set to 0.2.

Semantic verification loss. For present-query samples, the candidate-level verification label is $y_i^{\mathrm{ver}} = \mathrm{ind}[\mathrm{IoU}(B_i, B_c^{\mathrm{gt}}) \geq \delta]$. The verification loss is:

$$\mathcal{L}_{\mathrm{ver}} = \sum_{i \in \mathcal{I}_{\mathrm{pres}}} \mathrm{BCE}(V_i^{\mathrm{scv}}, y_i^{\mathrm{ver}}), \tag{20}$$

where $\mathcal{I}_{\mathrm{pres}}$ contains all candidate-query pairs from samples in which category c is present.

Absent-query rejection loss. For an absent query, a candidate is treated as an invalid hypothesis:

$$\mathcal{L}_{\mathrm{abs}} = \sum_{i \in \mathcal{I}_{\mathrm{abs}}} \mathrm{BCE}(V_i^{\mathrm{scv}}, 0), \tag{21}$$

where $\mathcal{I}_{\mathrm{abs}}$ contains all candidate-query pairs from samples in which category c is absent. Overall objective: $\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{con}} + \lambda_2 \mathcal{L}_{\mathrm{ver}} + \lambda_3 \mathcal{L}_{\mathrm{abs}}$.

4 Experiment

4.1 Datasets and Implementation Details

Datasets. We evaluate our method on two endoscopic surgical instrument segmentation benchmarks, EndoVis17 (Allan et al., 2019) and EndoVis18 (Allan et al., 2020). EndoVis17 contains 8 training and 2 validation sequences with 225 frames per sequence, while EndoVis18 contains 11 training and 4 validation sequences with 149 frames per sequence. Following prior work (Shvets et al., 2018), both datasets are preprocessed consistently with prior work. The evaluated instrument categories are Bipolar Forceps (BF), Prograsp Forceps (PF), Large Needle Driver (LND), Suction Instrument (SI), Vessel Sealer (VS), Clip Applier (CA), Grasping Retractor (GR), Monopolar Curved Scissors (MCS), and Ultrasound Probe (UP). EndoVis17 includes BF, PF, LND, VS, GR, MCS, and UP, whereas EndoVis18 includes BF, PF, LND, SI, CA, MCS, and UP.

Table 1: Overall comparison with state-of-the-art methods on EndoVis18 under weak-text prompts. The highest result is bolded, while the second highest result is underlined.

Dataset	Method	IoU	mcIoU	Instrument Categories						
				BF	PF	LND	SI	CA	MCS	UP
EndoVis18	SurgicalSAM	<u>80.33</u>	58.87	<u>83.66</u>	65.63	58.75	54.48	<u>39.78</u>	<u>88.56</u>	21.23
	TP-SIS(448)	**81.54**	<u>65.48</u>	81.53	<u>70.18</u>	<u>71.54</u>	**90.58**	21.46	65.57	<u>57.51</u>
	RoSIS	76.16	44.54	**84.11**	42.73	35.26	46.36	0.00	88.10	15.22
	CLIPSeg	69.86	39.70	67.25	39.59	36.72	47.27	2.92	79.96	4.22
	CRIS	72.29	46.04	73.58	58.20	47.64	72.14	4.56	45.99	20.18
	DaV (ours)	77.40	**82.39**	82.17	**72.43**	**83.87**	<u>79.00</u>	**79.53**	**94.80**	**84.90**

Metrics. We evaluate DaV at three levels: segmentation, target-confuser discrimination, and hypothesis verification. For segmentation, we report Intersection over Union (IoU) and mean class IoU (mcIoU), and use mcIoU as the primary metric to reduce the influence of class imbalance. For grounding-level evaluation, we report Target Selection Accuracy (TSA) and Confuser Error Rate (CER). TSA measures the percentage of present-query samples for which the top-ranked candidate overlaps the queried target by at least the matching threshold $\delta = 0.3$. CER is evaluated on valid target-confuser cases, where both the queried category and at least one confuser category are present and covered by the candidate set, and

measures the percentage of cases in which the highest-scoring confuser outranks the target. For hypothesis verification, we report Reject Accuracy and the area under the receiver operating characteristic curve (AUROC). Reject Accuracy measures the percentage of absent-query samples correctly predicted as empty, while AUROC evaluates how well the SCV verification score separates valid target hypotheses from invalid ones independently of the rejection threshold. Higher IoU, mcIoU, TSA, Reject Accuracy, and AUROC indicate better performance, whereas lower CER is better.

Implementation details. Our framework follows a text-to-box-to-mask pipeline: GroundingDINO (Liu et al., 2024) generates category-conditioned proposals, DaV performs discriminative discovery and selective verification, and the verified box is passed to a frozen SAM3 (Carion et al., 2025) for mask generation and temporal propagation. GroundingDINO keeps the top-5 proposals with a confidence threshold of 0.15. Each box is enlarged by 0.1, cropped, and resized to 224×224. Both visual and text features use a dimension of 256. We set the number of part tokens N_p to 5, latent attribute prototypes N_a to 16, and confuser categories N_c to 2. CPDG and SCV are jointly optimized with AdamW for 80 epochs using a learning rate and weight decay of 1×10^{-4}, with a batch size of 16. The loss weights λ_1, λ_2, and λ_3 are set to 1.0, 0.8, and 0.6, respectively, and the rejection threshold τ_{rej} is set to 0.5. Experiments are implemented in PyTorch with CUDA 12.0 on six NVIDIA GeForce RTX 3090 GPUs.

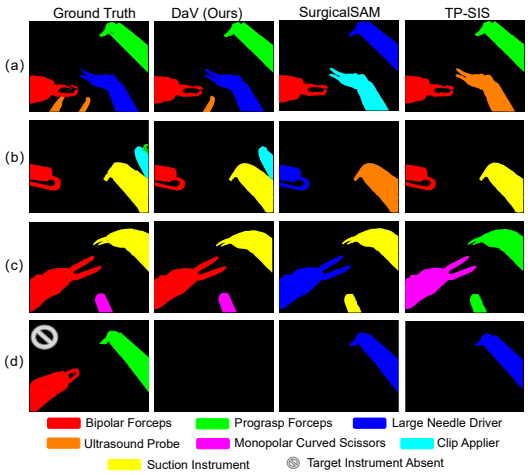


Figure 4: Qualitative comparison with other methods under weak-text prompts on EndoVis18.

4.2 Main Results on Weak-Text Fine-Grained Segmentation

Quantitative comparison. Tables 1 and 2 compare DaV with existing text-promptable surgical instrument segmentation methods on EndoVis18 and EndoVis17. Using mcIoU as the primary metric, DaV achieves 82.39% on EndoVis18 and 68.34% on EndoVis17. It also performs well across multiple instrument categories, including LND (83.87%), CA (79.53%), MCS (94.80%), and UP (84.90%) on EndoVis18, and BF (70.09%), MCS (94.99%), and UP (63.44%) on EndoVis17. These results show the downstream benefit of converting weak-text queries into verified spatial prompts. Sections 4.3 and 4.4 further analyze target-confuser discrimination and rejection reliability.

Qualitative comparison. Figure 4 illustrates two challenging weak-text scenarios. In rows (a)-(c), visually similar instruments appear together, and competing methods often activate non-target objects, whereas DaV better follows the queried category. In row (d), the target is absent; DaV outputs an empty mask while the compared methods produce false activations. Examples correspond to the two failure modes analyzed quantitatively in Sections 4.3 and 4.4: target-confuser ambiguity and absent-query rejection.

Table 2: Overall comparison with state-of-the-art methods on EndoVis17. The highest result is bolded, while the second highest result is underlined.

Dataset	Method	IoU	mcIoU	Instrument Categories						
				BF	PF	LND	VS	GR	MCS	UP
EndoVis17	SurgicalSAM	69.94	67.03	68.30	51.77	75.52	68.24	**57.63**	86.95	60.80
	TP-SIS(448)	**76.45**	54.78	69.57	**68.91**	**89.88**	**82.60**	0.00	72.53	0.00
	RoSIS	57.33	40.65	67.34	39.95	55.85	45.32	5.10	26.00	44.99
	CLIPSeg	65.02	33.42	51.29	42.27	49.56	30.12	9.96	30.69	20.05
	CRIS	67.83	38.21	54.87	50.21	68.33	50.12	0.00	43.97	0.00
	DaV (ours)	67.91	**68.34**	**70.09**	57.33	69.42	68.57	54.52	**94.99**	**63.44**

4.3 Target-Confuser Discrimination with CPDG

The central motivation of our work is that semantic similarity alone can be insufficient to distinguish the queried target from visually similar categories. As shown in Table 3, GroundingDINO achieves TSA values of 78.54% and 72.30% on EndoVis18 and EndoVis17, while its CER remains 51.80%

and 52.69%. We therefore directly evaluate whether CPDG improves target-confuser discrimination at the grounding stage. To isolate the effect of discriminative candidate scoring, we conduct an ablation study comparing GroundingDINO and CPDG. We evaluate candidate ranking using TSA and CER. This confirms that coarse semantic matching frequently ranks visually similar confusers above the queried target in the weak-text setting. CPDG increases TSA to 86.81% and 82.06% on EndoVis18 and EndoVis17, while reducing CER from 51.80% to 30.15% and 52.69% to 33.43%. The simultaneous improvement in overall target selection and reduction in target-confuser inversions shows that CPDG improves candidate discrimination before mask generation. These results directly support our motivation that fine-grained weak-text grounding benefits from explicitly modeling target-confuser differences rather than relying only on coarse semantic similarity. Figure 5 further shows that CPDG focuses on localized discriminative cues, whereas the baseline mainly responds to broadly shared structures.

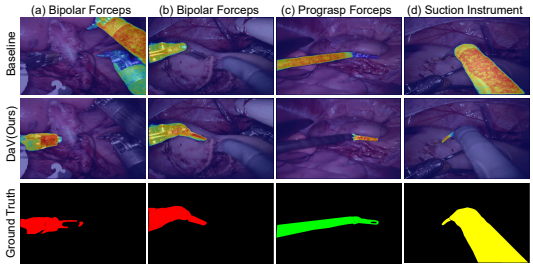


Figure 5: Visualization of discriminative evidence discovered by CPDG.

4.4 Hypothesis Verification with SCV

CPDG discovers target-discriminative evidence and uses it to identify the supported candidate hypothesis. However, discovery alone does not guarantee that the selected hypothesis is reliable, particularly when the queried category is absent. We evaluate whether SCV can verify the hypothesis and reject it when the available evidence is insufficient. We conduct an ablation study comparing SCV with a CPDG-only baseline that applies a rejection threshold of 0.5 without hypothesis verification. As shown in Table 3, SCV consistently outperforms the CPDG-only baseline, improving Reject Accuracy/AUROC from 84.09%/62.57% to 96.96%/80.43% on EndoVis18 and from 80.22%/55.91% to 94.13%/78.43% on EndoVis17. The higher AUROC indicates that the learned verification score better separates valid target hypotheses from unsupported candidates, while the improvement in Reject Accuracy shows that this separation translates into more reliable abstention when the queried category is absent. Figure 6 further illustrates that SCV suppresses insufficiently supported candidates and retains the hypothesis consistent with the query. Together with the absent-query example in Figure 4, these results demonstrate that SCV complements CPDG by verifying whether the discovered hypothesis is reliable enough to accept.

Table 3: Ablation of target-confuser discrimination and hypothesis verification on EndoVis17 and EndoVis18.

Dataset	Method	IoU ↑	mcIoU ↑	TSA ↑	CER ↓
EndoVis17	GroundingDINO	59.42	60.31	72.30	52.69
	CPDG	**62.04**	**63.17**	**82.06**	**33.43**
EndoVis18	GroundingDINO	70.49	76.64	78.54	51.80
	CPDG	**72.78**	**80.29**	**86.81**	**30.15**
Dataset	Method	IoU ↑	mcIoU ↑	Reject Acc. ↑	AUROC ↑
EndoVis17	CPDG only	62.04	63.17	80.22	55.91
	CPDG + SCV	**67.91**	**68.34**	**94.13**	**78.43**
EndoVis18	CPDG only	72.78	80.29	84.09	62.57
	CPDG + SCV	**77.40**	**82.39**	**96.96**	**80.43**

5 Conclusion

We study fine-grained video object segmentation from weak text prompts, where coarse semantic similarity can be insufficient to distinguish visually similar categories. To address this limitation, we propose DaV, a discover-and-verify framework that converts weak-text queries into reliable spatial prompts. CPDG explicitly models target-confuser differences to discover local category-discriminative evidence and identify the most supported candidate hypothesis, while SCV verifies whether this hypothesis has sufficient semantic and visual support and rejects it otherwise. Experiments on EndoVis17 and EndoVis18 demonstrate improved segmentation performance, while further analyses show better target-confuser discrimination and more reliable absent-query rejection. These results highlight the importance of combining discriminative evidence discovery with hypothesis verification for fine-grained segmentation under weak-text supervision.

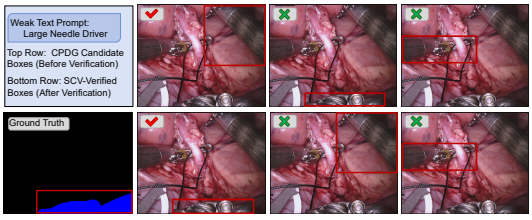


Figure 6: Qualitative illustration of hypothesis verification by SCV.

AI USE STATEMENT

In this work, generative AI tools (e.g., ChatGPT and Claude) assisted with code implementation, manuscript editing for readability, and reference formatting. All methodological decisions were made and independently validated by the authors. All AI-assisted code was reviewed, tested, and verified for correctness.

We did not use generative AI to generate synthetic datasets; develop theoretical or conceptual frameworks; formulate mathematical claims; propose or refine hypotheses; clean or reformat datasets; or interpret experimental results. As this paper contains no mathematical proofs, proof-related disclosure categories are not applicable.

All AI-assisted outputs were reviewed, AI-generated text was substantively revised as necessary, and citations were independently verified for authenticity and correct attribution. The authors take full responsibility for the final content, including all text, claims, code, and other AI-assisted artifacts.

REFERENCES

Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 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 preprint arXiv:2001.11190*, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Tae-Min Choi and Juyoun Park. Rethinking text-promptable surgical instrument segmentation with robust framework. *arXiv preprint arXiv:2411.12199*, 2024.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAM-WISE: Infusing wisdom in sam2 for text-driven video segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3395–3405. IEEE, 2025.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16301–16310, 2021.

Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Alexander Kirillov et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20009–20019. IEEE, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training

for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55. Springer, 2024.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7086–7096, 2022.

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 (ICLR)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021a.

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 (ICML)*, pp. 8748–8763. PMLR, 2021b.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 28085–28128, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Alexey A. Shvets, Alexander Rakhlin, Alexandr A. Kalinin, and Vladimir I. Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 624–628, 2018. doi: 10.1109/ICMLA.2018.00100.

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. doi: 10.1109/TMI.2024.3426953.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11686–11695, 2022.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10502–10511, 2019.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1307–1315, 2018.

Linfeng Yuan, Miaojing Shi, Zijie Yue, and Qijun Chen. Losh: Long-short text joint prediction network for referring video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14001–14010. IEEE, 2024.

Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgicalsam: Efficient class promptable surgical instrument segmentation. 2023.

Heli Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning discriminative object parts for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

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, pp. 28611–28623, 2023.

A APPENDIX

Loss-function Ablation To investigate the contribution of each training objective, we remove one loss term at a time while retaining the complete DaV architecture. As shown in Table 4, removing any objective reduces both mcIoU and Reject Accuracy on EndoVis17 and EndoVis18, indicating that the three objectives provide complementary supervision for segmentation and selective rejection.

Table 4: Ablation study of different loss terms on EndoVis17 and EndoVis18. The highest result is bolded, while the second-highest result is underlined.

Dataset	Loss	IoU ↑	mcIoU ↑	Reject Acc. ↑
EndoVis17	Ours w/o $\mathcal{L}_{\text{con}}$	64.86	64.16	<u>76.92</u>
	Ours w/o $\mathcal{L}_{\text{ver}}$	64.34	<u>64.29</u>	70.08
	Ours w/o $\mathcal{L}_{\text{abs}}$	<u>67.34</u>	64.17	72.50
	DaV (ours)	**67.91**	**68.34**	**94.13**
EndoVis18	Ours w/o $\mathcal{L}_{\text{con}}$	<u>76.28</u>	<u>80.89</u>	<u>78.53</u>
	Ours w/o $\mathcal{L}_{\text{ver}}$	72.82	80.27	76.36
	Ours w/o $\mathcal{L}_{\text{abs}}$	76.25	80.87	76.87
	DaV (ours)	**77.40**	**82.39**	**96.96**

Removing the target–confuser contrastive loss $\mathcal{L}_{\text{con}}$ decreases mcIoU from 68.34% to 64.16% on EndoVis17 and from 82.39% to 80.89% on EndoVis18. Reject Accuracy also decreases to 76.92% and 78.53%, respectively. These results are consistent with the role of contrastive supervision in learning discriminative evidence for CPDG, which also provides supporting information for subsequent verification.

Removing $\mathcal{L}_{\text{ver}}$ reduces mcIoU to 64.29% and 80.27%, while Reject Accuracy decreases to 70.08% and 76.36% on EndoVis17 and EndoVis18, respectively. This highlights the importance of candidate-level verification supervision in learning to distinguish valid target hypotheses from unsupported candidates. Without this supervision, absent-query rejection alone is insufficient to learn reliable hypothesis verification.

Removing $\mathcal{L}_{\text{abs}}$ decreases Reject Accuracy from 94.13% to 72.50% on EndoVis17 and from 96.96% to 76.87% on EndoVis18, demonstrating the importance of explicit absent-query supervision. The accompanying reductions in mcIoU further indicate that this objective contributes to the overall reliability of weak-text-prompted segmentation.

Overall, these results support the complementary roles of the three objectives: $\mathcal{L}_{\text{con}}$ promotes discriminative evidence discovery, $\mathcal{L}_{\text{ver}}$ supervises hypothesis validity, and $\mathcal{L}_{\text{abs}}$ explicitly supports absent-query rejection.

Table 5: Cross-dataset generalisation. *T*: training dataset; *V*: validation dataset; 18: EndoVis2018; 17: EndoVis2017. The highest result is bolded, while the second-highest result is underlined.

T	*V*	Method	Instrument Categories (IoU)				mcIoU
			BF	PF	LND	MCS	
18	17	MATIS Frame	45.57	32.62	44.98	58.84	45.50
		SurgicalSAM	<u>70.95</u>	35.21	45.46	<u>76.08</u>	<u>56.93</u>
		TP-SIS (896)	65.54	<u>58.19</u>	**84.01**	68.47	69.05
		DaV (ours)	**71.63**	**59.03**	<u>64.31</u>	**82.33**	**69.33**
17	18	MATIS Frame	<u>65.55</u>	13.89	38.25	**65.58**	45.81
		SurgicalSAM	44.50	27.17	<u>50.76</u>	<u>62.94</u>	<u>46.34</u>
		TP-SIS (896)	**65.81**	<u>56.12</u>	44.72	8.97	43.91
		DaV (ours)	30.47	**63.09**	**52.11**	54.18	**49.96**

Cross-Dataset Generalization. To assess generalization across datasets, we train DaV on EndoVis18 and evaluate it on EndoVis17, and vice versa. The evaluation covers four shared instrument categories: BF, PF, LND, and MCS.

As shown in Table 5, DaV achieves 69.33% mcIoU under EndoVis18-to-EndoVis17 transfer, compared with 69.05% for TP-SIS. In the reverse direction, DaV obtains 49.96% mcIoU, compared with

46.34% for SurgicalSAM. These results indicate that DaV retains competitive overall segmentation performance across datasets.

At the category level, DaV achieves the highest IoU on BF (71.63%), PF (59.03%), and MCS (82.33%) under EndoVis18-to-EndoVis17 transfer, while TP-SIS performs better on LND. Under EndoVis17-to-EndoVis18 transfer, DaV obtains the highest IoU on PF (63.09%) and LND (52.11%), but its performance on BF (30.47%) and MCS (54.18%) remains below that of competing methods.

These results suggest that the proposed framework can support fine-grained segmentation beyond the training dataset, while transfer performance remains category- and direction-dependent. In particular, the substantial variation on BF indicates that cross-dataset appearance differences remain a challenge for weak-text-guided target localization and segmentation.