# DECIDING IN THE DARK: MULTIMODAL LLMS AS EVIDENCE SELECTORS FOR FEW-SHOT ULTRASOUND VIDEO SEGMENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Few-shot segmentation of ultrasound sweeps depends on which annotated examples the segmenter sees and from which frame a mask is propagated, yet both are usually chosen by visual similarity, which is unreliable when speckle, shadowing and probe position change how the same anatomy looks. We present LucidSeg, a training-free framework in which a multimodal large language model (MLLM) makes discrete decisions about the evidence while frozen segmenters compute every mask. At Level 1, the MLLM chooses a small support set from retrieved candidates of other subjects, shown either as annotated exemplars or as annotations transferred onto the query, and a frozen in-context segmenter uses it; at Level 2, it chooses the starting frame of a sweep by judging the predicted outlines of candidate frames, and SAM2 propagates the mask from there. Because the MLLM never outputs a mask, box or point, its decisions serve any in-context segmenter. On four ultrasound datasets, a locally deployed open-weight MLLM matched a reranker trained on the target data and improved Dice over visual similarity by up to 0.11, with the largest gains where the target's appearance depends on anatomical side or probe position. On thyroid sweeps, the chosen starting frame came within 0.05 Dice of the best candidate; on spine sweeps over repeating vertebrae, starting from all candidates was better.

## 1 INTRODUCTION

Segmenting anatomical structures in ultrasound videos and sequences is an inherently ambiguous task. Speckle noise, acoustic shadowing, and view-dependent appearance often leave boundaries only partially visible, so even experienced annotators disagree on where a structure ends (Leclerc et al., 2019; Jiang et al., 2024). This ambiguity makes dense annotation costly and uneven in quality, and it varies considerably across anatomies: cardiac chambers are comparatively well delineated, whereas bony structures such as vertebrae appear as fragmented echoes above an acoustic shadow. For specialized tasks, such as scoliotic spine assessment, collecting a large annotated support set is rarely feasible. At the same time, the images themselves provide weak evidence: two frames that look alike may show different anatomy, and the same anatomy may look quite different across devices, patients, and probe positions.

Existing approaches follow two lines. Visual in-context methods (Butoi et al., 2023; Wong et al., 2025; Liu et al., 2024) and SAM2-based propagation (Ravi et al., 2025; Ma et al., 2025) choose supports or the starting frame by visual similarity or a fixed rule, which is unreliable when the target's appearance changes with probe position or anatomical side, or when noise dominates. Language-driven models (Rasaee et al., 2025; Xu et al., 2026; Liu et al., 2026b) map text to boxes or masks, but they rely on in-domain training and ask the language component for spatial output, where multimodal models are least reliable on ultrasound (Figure 1, top).

Both lines leave open which evidence the segmenter should see. Choosing the starting frame of propagation, or the exemplars that match the query in side, position or orientation, is a discrete judgment that draws on anatomy rather than pixel matching; in a pilot study, the starting frame alone changed the Dice of the same video from 0.00 to 0.76. Multimodal large language models (MLLMs) are poor at producing precise masks in ultrasound but can compare a few candidates and



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.

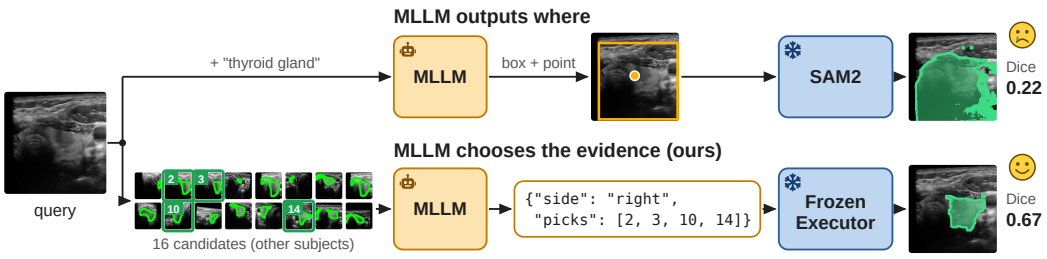



Figure 1: **Where the MLLM decides.** The same query (TG3K) and the same kind of MLLM. *Top:* as in reasoning segmentation, the MLLM localizes the target from its name and prompts SAM2 with a box and a point; on ultrasound the box covers most of the image (Dice 0.22). *Bottom:* in LucidSeg the MLLM only chooses four of the candidate annotated frames from other subjects, and a frozen in-context segmenter computes the mask (Dice 0.67). The MLLM produces no mask, box or point.

reason about anatomy, which suggests giving the MLLM the decision and a dedicated segmenter the geometry.

Based on this observation, we developed LucidSeg, a training-free framework in which an MLLM made discrete decisions about the evidence and frozen segmenters computed every mask (Figure 2). At Level 1, the MLLM chose four supports for the query from a short list of annotated frames retrieved without labels from other subjects. We presented the candidates either as annotated exemplars or with their annotations aligned to and drawn on the query. Where anatomy appears mirrored across scans, the MLLM reported the side of the query and a fixed rule oriented each support. At Level 2, the MLLM chose the starting frame of a sweep: among evenly spaced candidates segmented as at Level 1, it picked the one whose predicted outline fit best, and SAM2 propagated that mask through the sweep. Because the MLLM output no mask, box or point (Figure 1, bottom), its decisions transferred across segmenters without training. By default, we used locally deployed open-weight MLLMs, since clinical data often cannot be sent to third-party APIs. Our contributions are as follows:

- We formulated few-shot ultrasound segmentation as decisions about evidence, separating what an MLLM decides (which annotated examples to show and from which frame to start) from what frozen segmenters compute. The formulation required no training and applied unchanged to different in-context segmenters.
- We instantiated it as two decisions: support selection, with two ways of presenting the evidence to the MLLM (annotated exemplars, and annotations transferred onto the query), and the choice of the starting frame from predicted outlines.
- We evaluated LucidSeg on four ultrasound datasets (thyroid, prostate, heart and spine) against visual, text-driven and agent-based baselines. Without training and with a local open-weight model, it matched a reranker trained on the target data and improved on visual similarity where the target's appearance depends on anatomical side or probe position.

## 2 Related Work

### 2.1 Few-Shot and In-Context Segmentation

Classical few-shot medical segmentation matches support prototypes or descriptors against query features (Ouyang et al., 2020; Roy et al., 2020; Zhu et al., 2023; Cheng et al., 2024; Zhang et al., 2024; Tang et al., 2025). In-context models condition a single forward pass on a context set: UniverSeg (Butoi et al., 2023), MultiverSeg (Wong et al., 2025), Tyche (Rakic et al., 2024), Iris (Gao et al., 2025b) and K-Prism (Guo et al., 2026). Training-free pipelines turn supports into prompts or memory for SAM and SAM2 (Liu et al., 2024; Hu et al., 2024; Zu et al., 2025; He et al., 2025b; Xing et al., 2025), and ultrasound variants add learned prompts or temporal memory (Lin et al., 2024; Wahd et al., 2025; Deng et al., 2024; Kumaralingam et al., 2025). These methods differ in how they use the supports; which supports they receive, and from which frame a sequence is propagated, are treated as given.

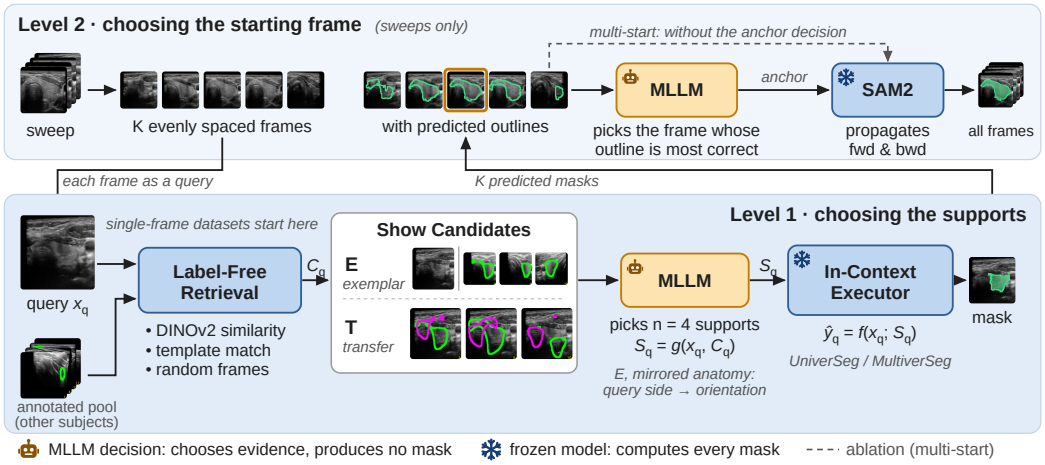



Figure 2: **Overview of LucidSeg.** The MLLM (amber) makes two discrete decisions about the evidence, and frozen models (blue) compute every mask. Level 1 (Section 3.2) chooses the supports of a query frame; Level 2 (Section 3.3) chooses the starting frame of a sweep. Dashed: the multi-start variant without the anchor decision.

### 2.2 Choosing the Evidence: Supports and Starting Frames

In-context performance depends strongly on which examples are supplied. Selection methods search for diverse contexts (Suo et al., 2024), rank candidate prompts (Xu et al., 2024), learn selection policies by reinforcement learning (Lee et al., 2026; Wu et al., 2024; Zhao et al., 2026), require semantic and mask agreement (Gao et al., 2025a), or optimize the support set (Uliel et al., 2026); the composition of the set matters more than its size (Gehad et al., 2026). For video, BubbleNets (Griffin & Corso, 2019) and interactive methods (Yin et al., 2021; Heo et al., 2021) choose the frame to annotate, and reasoning video segmentation lets an MLLM pick a keyframe in which to ground a referred object before propagation (Yan et al., 2024; Gong et al., 2025; Kao et al., 2026; Xu et al., 2025; Wen et al., 2026). None of these asks a model to judge from anatomy which annotated example or starting frame to trust.

### 2.3 Language and MLLMs in the Segmentation Pipeline

Language enters segmentation in two roles. In the first, it names the target class, either fused into the network (Karimi & Poullis, 2025; Wang et al., 2025; Chen et al., 2026a; Li et al., 2026) or used as a prompt for SAM-style segmenters (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026; Ren et al., 2024; Zhang et al., 2026; Ma et al., 2024; 2025; Aleem et al., 2024; Jiang et al., 2026a); for ultrasound, TextSAM-EUS (Spiegler et al., 2025), GroundingDINO-US-SAM (Rasaee et al., 2025) and UltraSAM3 (Xu et al., 2026) adapt these models on large ultrasound collections. In the second, an MLLM produces the spatial output: masks decoded from its tokens (Lai et al., 2024; Yuan et al., 2026), boxes and points for a frozen SAM from a reasoner trained with reinforcement learning (Liu et al., 2025; 2026b; Huang et al., 2025; Yan et al., 2026), or multi-turn interaction with the segmenter (Du et al., 2026; Zhu et al., 2025; He et al., 2025a; Liu et al., 2026a; Jiang et al., 2026b; Jin et al., 2026; Chen et al., 2026b). LucidSeg instead restricts the MLLM to discrete choices over annotated evidence and leaves every mask to frozen segmenters.

## 3 Method

LucidSeg separates what a multimodal LLM decides from what a segmenter computes. A frozen segmenter, the *executor*, produces every mask; the MLLM only makes discrete choices about the evidence the executor receives. We used two such decisions: which annotated frames from other subjects serve as supports for a query frame (Level 1, Section 3.2), and, for a sweep, from which frame the segmentation is propagated (Level 2, Section 3.3). Neither the MLLM nor the executor was trained or fine-tuned. Figure 2 gives an overview.

### 3.1 Problem Setting

A query is an unlabelled ultrasound frame $x_q$. An annotated pool $\mathcal{P} = \{(x_i, y_i)\}$ holds frames and masks from other subjects; only the pool frames from units other than that of the query are allowed, which gives $\mathcal{P}_q \subseteq \mathcal{P}$. An in-context executor $f$ segments the query from a support set $S \subseteq \mathcal{P}_q$ of fixed size $n$: $\hat{y}_q = f(x_q; S)$. The quality of $\hat{y}_q$ depends strongly on $S$: on TG3K, the best set of four supports in the pool reaches 0.87 Dice with UniverSeg, whereas a random set reaches 0.38 (Table 1). Choosing $S$ without the query's mask is the problem of Level 1. We treated $f$ as a black box, so that the same decision can serve any in-context executor.

### 3.2 Level 1: Choosing the Supports

**Candidate proposal.** An MLLM cannot inspect a pool of hundreds of frames in one call, so label-free retrieval first proposed a short candidate list $\mathcal{C}_q \subseteq \mathcal{P}_q$. It was the union of three lists of $k$ frames each: the frames most similar to the query by DINOv2 CLS embedding (Oquab et al., 2024), the frames whose annotated region best matches the query by normalized cross-correlation, and frames drawn at random. The two retrieval criteria make the list likely to contain good supports, and the random part keeps it from depending only on visual similarity, which is a weak guide when the appearance of the target changes with probe position or anatomical side, or when noise dominates the image. The list was shuffled before it was shown to the model, so that its order carried no information about the retrieval scores.

**Presenting the evidence.** We compared two ways of showing a candidate $c = (x_c, y_c) \in \mathcal{C}_q$ to the MLLM. In the *exemplar* mode (LucidSeg$_{\mathrm{E}}$), each candidate appeared as its own image $x_c$ with the outline of $y_c$ drawn on it, next to the query. The model compared the query with annotated examples, as a clinician would compare a scan with reference images. In the *transfer* mode (LucidSeg$_{\mathrm{T}}$), each candidate's annotation was carried onto the query. A similarity transform without rotation, $T_c$ (translation and isotropic scale), was fitted by matching the annotated region of $x_c$ against $x_q$, and the outline of $T_c(y_c)$ was drawn on a copy of the query. The model then judged directly whether a transferred outline lies on the target structure of the query. A misplaced outline is itself evidence against the candidate. In both modes the candidates were numbered, and no text description of the candidates was given.

**Decision.** In one call, the MLLM received the query and the rendered candidates and returned the indices of $n$ distinct candidates, $S_q = g(x_q, \mathcal{C}_q)$, as structured output. In the exemplar mode it first stated an anatomical reading of the query: the side of the thyroid gland in the image, or the lateral position of a prostate slice. Stating this reading first ties the choice to what the model sees in the query, and on TG3K it also sets the orientation of the supports; Appendix L reports how accurately the side is read and what its errors cost.

**Orientation.** Some anatomies appear in mirrored form across scans: transverse scans of the left and right thyroid lobe are mirror images of each other, and a support from the opposite side transfers poorly. On such data each candidate could also be used mirrored. In the exemplar mode, the orientation followed from the anatomical reading by a fixed *side rule*: a chosen support was mirrored when the side of its own mask, taken from the horizontal position of the mask centroid, was opposite to the query side the model had reported. The model thus judges only the query, not pairs of mirrored images. Appendix H compares this rule with letting the model decide the orientation. In the transfer mode, each tile showed the candidate's outline in both orientations in two colours, and the model's choice of colour set the orientation. Mirroring was used only where the anatomy calls for it (TG3K in our experiments).

**Execution.** The chosen supports, as original images and masks and mirrored where required, formed the context set of the executor, which segmented the query: $\hat{y}_q = f(x_q; S_q)$. The MLLM produced no mask, point or box, and the executor never saw the rendered tiles. The same decision can therefore be passed to different executors; we used UniverSeg (Butoi et al., 2023) and MultiverSeg (Wong et al., 2025).

### 3.3 LEVEL 2: CHOOSING THE STARTING FRAME

A sweep is a sequence of unlabelled frames $V = (x_1, \ldots, x_T)$ of one subject, and the goal is a mask for every frame. SAM2 (Ravi et al., 2025) propagates a mask given on one frame, the *anchor*, forwards and backwards through the sequence. The anchor matters: near the ends of a sweep the target is often small, partly outside the image or noisy, and a poor anchor mask spreads its errors through the whole sweep. Level 2 lets the MLLM choose the anchor.

**Candidate anchors.** We took $K$ candidate frames evenly spaced over the sweep, including its first and last frame. Each candidate $x_t$ was segmented as at Level 1: supports were chosen for it from the pool, and the executor produced a predicted mask $\hat{y}_t = f(x_t; S_t)$. In the full system the MLLM chose these supports; to isolate the anchor decision, we also fixed them to the DINOv2 selection.

**Decision.** Each candidate frame was shown with the outline of its predicted mask $\hat{y}_t$, never the ground truth, and the $K$ frames were numbered in time order. In one call the MLLM chose the frame whose outline was most correct: it covers the whole target, follows its boundary, has a typical shape and includes no other structure. The prompt noted that the target is often small or incomplete near the ends of the sweep. The anchor decision thus judged the result of the support decision rather than the raw appearance of the frames. Human annotators likewise start propagation from a frame whose segmentation is complete, and in video object segmentation the starting frame strongly affects the quality over the whole video (Griffin & Corso, 2019; Yin et al., 2021).

**Propagation.** The predicted mask of the chosen anchor $a$ initialised SAM2, which propagated it to the start and to the end of the sweep and returned a mask for every frame. A variant that chooses the anchor from the raw frames, without the predicted outlines, is compared in Appendix I.

**Without the anchor decision.** Because the $K$ candidates are already segmented, their predicted masks can also be given to SAM2 together as conditioning masks, without choosing an anchor (*multi-start*). This variant keeps the support decision and removes the anchor decision, so comparing it with the chosen anchor measures what the anchor decision adds. We report it with all $K$ candidates and without the first and last frame.

## 4 EXPERIMENTAL SETUP

### 4.1 EVALUATION PROTOCOL

Level 1 was evaluated on all four datasets. Each test query was a single frame, every method drew on the same pool of annotated frames from other subjects with the same budget of $n{=}4$ supports, and the query list was frozen before any run. Level 2 needs sweeps in which every frame is annotated, so it was evaluated on TG3K (the 13 test scans) and USST (6 of the 18 test patients, evenly spaced). On USST, each sweep joins the contiguous vertebral segments of one patient (Appendix A). CardiacUDA has five annotated frames per video and was evaluated at Level 1 only. The ground truth of a query was used only for scoring and for the oracle reference rows.

### 4.2 DATASETS

We used three public datasets and one private dataset, each split leave-one-unit-out by patient or video; the pool held 8 annotated frames of every other unit. **TG3K** (Gong et al., 2021): transverse thyroid scans, 260 test queries from 13 scans. **MicroSeg** (Jiang et al., 2024): sagittal micro-ultrasound of the prostate, 2,621 annotated slices of 75 patients, 355 test queries. **CardiacUDA** (Yang et al., 2023): apical four-chamber echocardiography from two hospitals, with the pool from one and 100 test queries from the other. **USST**: a private, de-identified dataset of thoracic spine sweeps, 2,980 annotated frames of 21 patients, 250 test queries (Appendix C). Supports could be mirrored only on TG3K, and each cardiac chamber was segmented as a binary task. Details are in Appendix B.

### 4.3 Baselines and Metrics

We compared against three groups of methods, all with official or default settings and at most one documented adjustment (Appendix G). *Visual-based:* UniverSeg (Butoi et al., 2023) and MultiverSeg (Wong et al., 2025) with $n$ random supports, and visual criteria that choose the $n$ supports for the same executors from the same pool as LucidSeg: DINOv2 similarity (Oquab et al., 2024), template matching by normalized cross-correlation (NCC), and a reranker trained on the target dataset with cross-validation by unit. Matcher (Liu et al., 2024) is reported on TG3K in Appendix J. *Text-driven:* GroundingDINO-US-SAM (Rasaee et al., 2025) and UltraSAM3 (Xu et al., 2026), prompted with the class names they were trained on and given no supports; their released weights were trained on TG3K and MicroSeg (UltraSAM3 also on CardiacUDA), so we mark those results as in-domain; USST is unseen by both. *Agent-based:* VisionReasoner (Liu et al., 2026b), a reasoning MLLM that prompts a frozen SAM2 with a box and a point, run zero-shot. We report Dice at the native resolution of the query (averaged over the four chambers on CardiacUDA) as the mean over test queries, with 95% confidence intervals from a cluster bootstrap over units and paired Wilcoxon signed-rank tests; at Level 2, we report the mean Dice over all frames of a sweep. IoU, HD95 and the reference rows are defined in Appendix D.

### 4.4 Implementation Details

The candidate list combined $k{=}6$ frames from each of the three sources (14–18 unique candidates), rendered at 512 px, and the model chose $n{=}4$ supports; at Level 2 it chose among $K{=}8$ evenly spaced frames, each shown with the executor's predicted outline, and SAM2.1-Large propagated the anchor mask. The main setting used locally deployed open-weight models, Qwen3.5-9B (Qwen Team, 2026) and Gemma-4-31B (Gemma Team, Google DeepMind, 2026), at temperature 0 with thinking disabled; the closed models GPT-5.5 and Claude Sonnet 5 received only de-identified images. Level 2 used Gemma-4-31B and, for the full system, also GPT-5.5. Further details, the prompts and the de-identification procedure are in Appendices F and E.

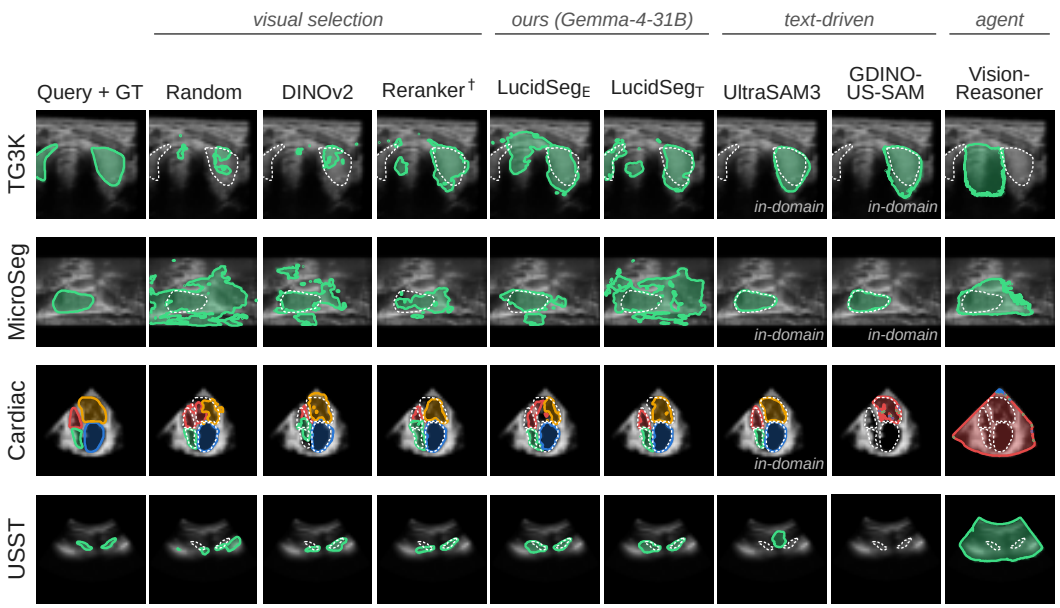
## 5 Main Results

### 5.1 Level 1: Support Selection

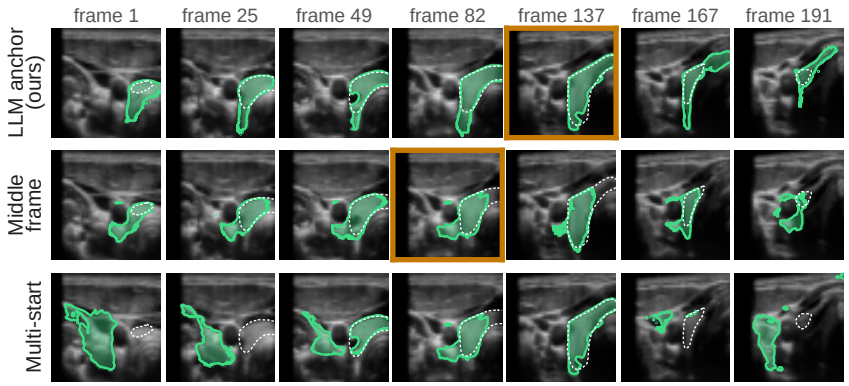
Table 1 shows the Dice of all methods at Level 1, and Figure 3a shows one query per dataset, chosen by a fixed rule: among the queries where LucidSeg$_{\mathrm{E}}$ beats DINOv2 similarity, the one whose LucidSeg$_{\mathrm{E}}$ Dice is closest to its mean over the dataset. All support-based rows use the same pool, queries and budget of four supports; on TG3K, the visual policies may pick mirrored copies, and LucidSeg mirrors by the side rule. GPT-5.5 ran without reasoning.

**Which supports are used matters.** On every dataset, the Dice reached from the pool depended strongly on the supports: from random supports to the oracle set, UniverSeg rose from 0.38 to 0.87 on TG3K and from 0.55 to 0.88 on USST. Visual similarity recovered little of this gap where the appearance of the target changes with anatomical side or probe position. On TG3K, DINOv2 similarity was no better than random supports ($+0.017$, $p{=}0.14$), and template matching was worse. On MicroSeg, in contrast, DINOv2 similarity was the strongest visual policy, 0.14 above random supports.

**Choosing supports with an MLLM.** On TG3K, where the lobe side must be judged, every LucidSeg configuration with Gemma-4-31B or GPT-5.5 improved on DINOv2 similarity by 0.08–0.11 Dice with UniverSeg ($p{<}10^{-6}$), and the transfer mode with Gemma-4-31B also improved on the reranker trained on TG3K ($+0.036$, $p{=}0.012$). On USST, the transfer mode improved on DINOv2 similarity ($+0.034$ with Gemma-4-31B, $p{=}0.005$) and matched the reranker ($-0.012$, $p{=}0.63$), whereas the exemplar mode of Gemma-4-31B stayed at the level of DINOv2 ($+0.006$, $p{=}0.58$). On CardiacUDA all support policies lay within 0.04 of each other, and the transfer mode was best with both executors. On MicroSeg, the MLLM matched the reranker but remained below DINOv2 similarity (with Gemma-4-31B, $-0.023$ with UniverSeg, $p{=}0.19$, and $-0.031$ with MultiverSeg): the prostate has a single shape in the sagittal view and differs mainly in size and position, so retrieval by appearance already finds suitable supports and leaves no anatomical judgment to make.




(a)




(b)

Figure 3: **Segmentations at both levels.** Predictions in green, ground truth as a white dashed outline (on CardiacUDA, one colour per chamber). **(a)** Level 1 with UniverSeg on one typical test query per dataset; LucidSeg uses Gemma-4-31B. †Trained on the target dataset. **(b)** Level 2 on a TG3K test sweep of median difficulty (full system, Gemma-4-31B, MultiverSeg); the orange border marks the frame from which the row is propagated. More visualization results are provided in Appendix O.

The same decisions transferred between executors: on TG3K the best LucidSeg row exceeded the best visual-based row, including the reranker, by 0.036 with UniverSeg and 0.032 with MultiverSeg. Among the four MLLMs, Gemma-4-31B and GPT-5.5 were the strongest (Appendix K); the paired differences with confidence intervals are shown in Figure 5 in Appendix N.

**Baselines without supports.** UltraSAM3 was the strongest on its training datasets (in-domain), but on USST, which neither had seen, the two reached only 0.111 and 0.032 Dice. VisionReasoner, which asks the MLLM to localize the target itself, reached 0.14–0.51. Where no in-domain model exists, as for the spine, the zero-shot alternatives fail, while choosing among annotated examples from other subjects still gives usable segmentations.

Table 1: **Level 1: support selection and baselines.** Dice, mean over test queries, with UniverSeg (UvS) and MultiverSeg (MvS). †Trained on the target dataset. Grey: in-domain (trained on this dataset). Bold: best support-based result per column. –: not run or not available.

| | TG3K | | MicroSeg | | CardiacUDA | | USST | |
|---|---|---|---|---|---|---|---|---|
| Method | UvS | MvS | UvS | MvS | UvS | MvS | UvS | MvS |
| *Visual-based* | | | | | | | | |
| Random supports (Butoi et al., 2023; Wong et al., 2025) | .384 | .547 | .582 | .702 | .663 | .757 | .545 | .604 |
| DINOv2 similarity | .401 | .544 | **.718** | **.802** | .682 | .786 | .570 | .646 |
| Template matching (NCC) | .307 | .480 | .264 | .430 | .678 | .770 | .582 | .614 |
| Reranker† | .472 | .603 | .687 | .763§ | .688 | .780 | **.616** | .678 |
| *Text-driven and agent-based (no supports)* | | | | | | | | |
| GroundingDINO-US-SAM (Rasaee et al., 2025) | .784 | | .699 | | .210 | | .111 | |
| UltraSAM3 (Xu et al., 2026) | .852 | | .794 | | .836 | | .032 | |
| VisionReasoner (Liu et al., 2026b) | .251 | | .507 | | .169 | | .139 | |
| *Ours (no training)* | | | | | | | | |
| LucidSeg$_{\mathrm{E}}$, Gemma-4-31B | .484 | .611 | .694 | .771 | .665 | .767 | .576 | .650 |
| LucidSeg$_{\mathrm{T}}$, Gemma-4-31B | **.508** | **.635** | .674 | .748 | **.703** | .784 | .604 | **.679** |
| LucidSeg$_{\mathrm{E}}$, GPT-5.5 | .498 | .618 | .698 | .754 | .687 | .776 | .593 | .666 |
| LucidSeg$_{\mathrm{T}}$, GPT-5.5 | .497 | .633 | .663 | .738 | **.703** | **.787** | .593 | .654 |
| $\Delta$ vs. best visual-based | $+.036$ | $+.032$ | $-.020$ | $-.031$ | $+.015$ | $+.001$ | $-.012$ | $+.001$ |
| *References (use ground truth; not deployable)* | | | | | | | | |
| Union-oracle (best set in list) | .779 | .860 | .875 | – | .824 | .869 | .789 | .838 |
| Oracle (whole pool) | .868 | .900 | .930 | – | .873 | .898 | .880 | .908 |
| Same-video supports | .728 | .786 | .632 | – | – | – | – | – |

§MicroSeg, MvS: the reranker trained on UniverSeg Dice (cross-executor); the MvS column re-runs the sets chosen for UvS.

### 5.2 Level 2: Anchor Selection and the Full System

Table 2 shows the Dice over whole sweeps at Level 2, and Figure 3b shows seven frames of one TG3K sweep propagated from the LLM anchor, from the middle candidate and from all candidates at once (the median sweep among those where the LLM did not choose the middle candidate). The anchor of the DINOv2 chain was chosen by Gemma-4-31B.

**The starting frame matters.** Starting from the first frame of the sweep was the worst policy on TG3K, and a random candidate lost 0.14–0.21 Dice against the LLM anchor on TG3K ($p \leq 0.0007$), so where propagation starts decides much of the final quality. Near the ends of a sweep the target is small or incomplete, and a mask started there spreads its errors (Figure 3b).

**Single anchor or several starts.** On TG3K, the anchor chosen from the predicted outlines reached 0.780–0.785 in all three chains, within 0.05 of the best candidate, and beat multi-start by 0.057–0.123 ($p \leq 0.027$); with GPT-5.5 it also beat the middle frame ($+0.051$, $p=0.012$), while Gemma-4-31B mostly chose the middle candidate itself (9 of 13 sweeps). The predicted outlines were essential: from the raw frames, the MLLM never clearly beat the middle frame (Appendix I). On USST, whose sweeps pass several similar vertebrae, multi-start was best in every chain and exceeded the best single anchor, although six sweeps allowed no significant difference. One structure per sweep thus favours a well-chosen anchor, and repeating structures favour many starts.

## 6 Ablation Studies

Table 3a shows the effect of the query frame, and Table 3b the effect of text descriptions of the candidates.

**The query frame.** With GPT-5.5, showing the query improved the decision on all four datasets (by 0.017–0.041 with UniverSeg). With Gemma-4-31B, the query mattered most on MicroSeg ($+0.056$), whereas on the other datasets the decision without the query was as good; on TG3K without the query, Gemma-4-31B chose clear, representative examples and still improved on DINOv2 similarity by 0.08. The stronger model thus relied more on the query.

Table 2: **Level 2: anchor selection and the full system.** Dice over all frames of a sweep after SAM2 propagation (executor MultiverSeg, $K=8$). Rows: how the anchor is chosen; columns: who chooses the supports. Shaded: the full system. Bold: best result per column without ground truth. n/a: not run.

| | TG3K (13 sweeps): supports by | | | USST (6 sweeps): supports by | | |
|---|---|---|---|---|---|---|
| Anchor policy | DINOv2 | Gemma | GPT-5.5 | DINOv2 | Gemma | GPT-5.5 |
| First frame | .427 | .587 | .488 | .484 | .588 | .575 |
| Middle frame | .745 | .772 | .735 | .623 | .640 | .633 |
| Max. image entropy | .641 | .711 | .743 | .485 | .550 | .530 |
| DINOv2 medoid | .678 | .700 | .727 | .610 | .637 | .549 |
| Executor confidence | .737 | **.802** | .754 | .589 | .630 | .622 |
| Random frame (expected) | .566 | .639 | .612 | .576 | .609 | .564 |
| LLM anchor: Gemma-4-31B | **.780** | .781 | n/a | .678 | .665 | n/a |
| LLM anchor: GPT-5.5 | n/a | n/a | **.785** | n/a | n/a | .632 |
| *Multi-start: all candidate predictions condition SAM2 (no anchor decision)* | | | | | | |
| All $K$ frames | .657 | .693 | .709 | .712 | **.726** | **.724** |
| Inner $K-2$ frames | .684 | .694 | .728 | **.713** | .725 | .709 |
| *References (use ground truth; not deployable)* | | | | | | |
| Anchor oracle (best of $K$) | .828 | .825 | .833 | .706 | .706 | .695 |
| GT anchor mask, best of $K$ | n/a | .867 | n/a | n/a | .755 | n/a |

Table 3: **What the decision sees** (LucidSeg$_{\mathrm{E}}$, Dice UniverSeg / MultiverSeg). **(a)** Query frame shown or withheld. **(b)** Text descriptions of the candidates (TG3K).

| | Query | TG3K | MicroSeg | CardiacUDA | USST |
|---|---|---|---|---|---|
| Gemma-4-31B | shown | .484 / .611 | .694 / .771 | .665 / .767 | .576 / .650 |
| | withheld | .483 / .622 | .638 / .723 | .682 / .778 | .580 / .675 |
| GPT-5.5 | shown | .498 / .618 | .698 / .754 | .687 / .776 | .593 / .666 |
| | withheld | .468 / .601 | .657 / .732 | .670 / .766 | .572 / .673 |

(a)

| Candidates as | Qwen3.5-9B | Gemma-4-31B |
|---|---|---|
| images | .447 / .591 | .472 / .593 |
| images + text | .437 / .567 | .468 / .589 |
| text only | .389 / .559 | .453 / .603 |

(b)

**Descriptions and robustness.** Adding text descriptions of the candidates, written by the same model, did not help on TG3K (Table 3b, with the models' own mirror flags), and descriptions alone were worse for Qwen3.5-9B; we therefore show images only. The order of the candidates mattered little for Qwen3.5-9B in the exemplar mode: over five orders on TG3K, its Dice varied with a standard deviation of 0.003 (UniverSeg) and 0.008 (MultiverSeg). For Gemma-4-31B in the transfer mode, one other order lowered Dice from 0.508 to 0.484, so differences of about 0.02 between configurations are within this variation.

**Anatomical reading.** On TG3K, Gemma-4-31B and GPT-5.5 read the query side correctly in 68% and 77% of the clear cases; the side rule built on this reading beat their own mirror decisions, and the ground-truth side would add 0.01–0.04 (Appendices L and H).

## 7 Conclusion

We showed that a multimodal LLM can guide few-shot ultrasound segmentation without producing any mask, by deciding which annotated examples a frozen segmenter sees and from which frame a sweep is propagated. Without training, these decisions improved on visual similarity where the target's appearance depends on anatomical side or probe position, matched a reranker trained on the target data with a local open-weight model, and brought the starting frame of thyroid sweeps close to the best candidate. Where appearance retrieval already suffices, or where a sweep passes repeating structures, the decision added little, which marks the scope of the approach and points to deciding how many starting frames to use as a next step.

### AI USE STATEMENT

In this work, we designed the experiments and structured the paper ourselves, and used AI tools to polish and edit the text of the paper, to write parts of the experimental code, and to carry out parts of the experiments under our specification. We reviewed all AI-assisted work: the authors checked AI-generated code for correctness before use and verified that all reported numbers were produced by the logged runs. The multimodal language models evaluated as part of our method are objects of study. We take responsibility for the final content of this work.

### ETHICS STATEMENT

The private USST dataset was collected with the approval of the institutional ethics board, and all images were de-identified before use. All other datasets (TG3K, MicroSeg and CardiacUDA) are publicly available, and we used them under their respective licenses. Before any image was sent to a closed-source model, we masked everything outside the imaged sector to remove burned-in text, and prompts contained no patient identifiers. USST images were sent only to providers whose use for research was approved for this dataset. Our method is a research prototype and was not validated for clinical use; its outputs should not inform diagnosis or treatment without further clinical evaluation.

### REPRODUCIBILITY STATEMENT

We provide the essential components for reproducing our results in the supplementary material, including detailed settings, more comparisons and ablation studies. We will release the code upon acceptance, together with the query lists and splits for the public datasets; the USST data will be released in accordance with the terms of its ethics approval.

## REFERENCES

Sidra Aleem, Fangyijie Wang, Mayug Maniparambil, Eric Arazo, Julia Dietlmeier, Kathleen Curran, Noel E. O'Connor, and Suzanne Little. Test-time adaptation with SaLIP: A cascade of SAM and CLIP for zero-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5184–5193, 2024.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. UniverSeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21438–21451, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026.

H. Chen, A. Xu, L. Zhang, M. Ouyang, Z. Zhu, and J. Wang. Text-guided few-shot liver and tumor segmentation. *Frontiers in Digital Health*, 8, 2026a. doi: 10.3389/fdgth.2026.1823870.

Zhigang Chen, Xiawu Zheng, and Rongrong Ji. Toward robust in-context segmentation via concept guidance. *arXiv preprint arXiv:2606.28149*, 2026b.

Ziming Cheng, Shidong Wang, Tong Xin, Tao Zhou, Haofeng Zhang, and Ling Shao. Few-shot medical image segmentation via generating multiple representative descriptors. *IEEE Transactions on Medical Imaging*, 43(6):2202–2214, 2024.

Xiaolong Deng, Huisi Wu, Runhao Zeng, and Jing Qin. MemSAM: Taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9622–9631, 2024.

Tianyuan Du, Haopeng Li, Zhen Fan, Jiarui Zhang, Panwang Pan, and Yang Zhang. SAM-Veteran: An MLLM-based human-like SAM agent for reasoning segmentation. In *International Conference on Learning Representations*, 2026.

Jun Gao, Qicheng Lao, Qingbo Kang, Paul Liu, Chenlin Du, Kang Li, and Le Zhang. Boosting your context by dual similarity checkup for in-context learning medical image segmentation. *IEEE Transactions on Medical Imaging*, 44:310–319, 2025a. doi: 10.1109/TMI.2024.3440311.

Yunhe Gao, Di Liu, Zhuowei Li, Yunsheng Li, Dongdong Chen, Mu Zhou, and Dimitris N Metaxas. Show and segment: Universal medical image segmentation via in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20830–20840, 2025b.

Youssef Gehad, Emmanuel Zerefa, Krish Kabra, and Guha Balakrishnan. Context matters: Support set selection and failure detection for in-context medical image segmentation. *arXiv preprint arXiv:2608.05333*, 2026.

Gemma Team, Google DeepMind. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 257–261, 2021. doi: 10.1109/ISBI48211.2021.9434087.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29183–29192, 2025.

Brent A. Griffin and Jason J. Corso. BubbleNets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Bangwei Guo, Yunhe Gao, Meng Ye, Difei Gu, Yang Zhou, Leon Axel, and Dimitris Metaxas. K-Prism: A knowledge-guided and prompt integrated universal medical image segmentation model. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xingqi He, Yujie Zhang, Shuyong Gao, Wenjie Li, Lingyi Hong, Mingxi Chen, Kaixun Jiang, Jiyuan Fu, and Wenqiang Zhang. RSAgent: Learning to reason and act for text-guided segmentation via multi-turn tool invocations. *arXiv preprint arXiv:2512.24023*, 2025a.

Xingxin He, Yifan Hu, Zhaoye Zhou, Mohamed Jarraya, and Fang Liu. Few-shot adaptation of training-free foundation model for 3d medical image segmentation. *arXiv preprint arXiv:2501.09138*, 2025b.

Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Guided interactive video object segmentation using reliability-based attention maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7322–7330, 2021.

Jiesi Hu, Yang Shang, Yanwu Yang, Xutao Guo, Hanyang Peng, and Ting Ma. ICL-SAM: Synergizing in-context learning model and SAM in medical image segmentation. In *Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 641–656. PMLR, 2024.

Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. SAM-R1: Leveraging SAM for reward feedback in multimodal segmentation via reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

Chongcong Jiang, Tianxingjian Ding, Chuhan Song, Jiachen Tu, Ziyang Yan, Yihua Shao, Zhenyi Wang, Yuzhang Shang, Tianyu Han, and Yu Tian. Medical SAM3: A foundation model for universal prompt-driven medical image segmentation. *arXiv preprint arXiv:2601.10880*, 2026a.

Hongxu Jiang, Muhammad Imran, Preethika Muralidharan, Anjali Patel, Jake Pensa, Muxuan Liang, Tarik Benidir, Joseph R. Grajo, Jason P. Joseph, Russell Terry, John Michael DiBianco, Li-Ming Su, Yuyin Zhou, Wayne G. Brisbane, and Wei Shao. MicroSegNet: A deep learning approach for prostate segmentation on micro-ultrasound images. *Computerized Medical Imaging and Graphics*, 112:102326, 2024.

Yankai Jiang, Qiaoru Li, Binlu Xu, Haoran Sun, Chao Ding, Junting Dong, Yuxiang Cai, Xuhong Zhang, and Jianwei Yin. IBISAgent: Reinforcing pixel-level visual reasoning in MLLMs for universal biomedical object referring and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Woojeong Jin, Jaeho Lee, Heeseong Shin, Seungho Jang, Junhwan Heo, and Seungryong Kim. AgentRVOS: Reasoning over object tracks for zero-shot referring video object segmentation. *arXiv preprint arXiv:2603.23489*, 2026.

Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. CoT-RVS: Zero-shot chain-of-thought reasoning segmentation for videos. In *The Fourteenth International Conference on Learning Representations*, 2026.

Amin Karimi and Charalambos Poullis. DSV-LFS: Unifying LLM-driven semantic cues with visual features for robust few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4584–4594, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Logiraj Kumaralingam, Anparasy Sivaanpu, Manh-Hai Hoang, Javaneh Alavi, Kim-Cuong T Nguyen, Kumaradevan Punithakumar, Edmond HM Lou, Paul Major, and Lawrence H Le. OralSAM: One-shot segmentation for intraoral ultrasound videos with adaptive feature correlation and self-prompting strategy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 390–399. Springer, 2025. doi: 10.1007/978-3-032-04981-0_37.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Lovstakken, and Olivier Bernard. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Transactions on Medical Imaging*, 38(9):2198–2210, 2019.

Eugene Lee, Yu-Chi Lin, and Jiajie Diao. Learning to select visual in-context demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9455–9465, 2026.

Meihua Li, Yang Zhang, Weizhao He, Hu Qu, and Yisong Li. SD-FSMIS: Adapting stable diffusion for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29979–29989, 2026.

Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 24–34. Springer, 2024.

Shengyuan Liu, Liuxin Bao, Qi Yang, Wanting Geng, Boyun Zheng, Chenxin Li, Wenting Chen, Houwen Peng, and Yixuan Yuan. MedSAM-Agent: Empowering interactive medical image segmentation with multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2602.03320*, 2026a.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *The Twelfth International Conference on Learning Representations*, 2024.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. VisionReasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. MedSAM2: Segment anything in 3D medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pp. 762–780. Springer, 2020.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, 2026. Model card: `https://huggingface.co/Qwen/Qwen3.5-9B`.

Marianne Rakic, Hallee E. Wong, Jose Javier Gonzalez Ortiz, Beth A. Cimini, John V. Guttag, and Adrian V. Dalca. Tyche: Stochastic in-context learning for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11159–11173, 2024.

Hamza Rasaee, Taha Koleilat, and Hassan Rivaz. Grounding DINO-US-SAM: Text-prompted multiorgan segmentation in ultrasound with LoRA-tuned vision-language models. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 72(10):1414–1425, 2025. doi: 10.1109/TUFFC.2025.3605285.

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 *The Thirteenth International Conference on Learning Representations*, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. 'squeeze & excite' guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020.

Pascal Spiegler, Taha Koleilat, Arash Harirpoush, Corey S Miller, Hassan Rivaz, Marta Kersten-Oertel, and Yiming Xiao. TextSAM-EUS: Text prompt learning for SAM to accurately segment pancreatic tumor in endoscopic ultrasound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 959–968, 2025.

Wei Suo, Lanqing Lai, Mengyang Sun, Hanwang Zhang, Peng Wang, and Yanning Zhang. Rethinking and improving visual prompt selection for in-context learning segmentation. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72952-2_2.

Song Tang, Shaxu Yan, Xiaozhi Qi, Jianxin Gao, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Few-shot medical image segmentation with high-fidelity prototypes. *Medical Image Analysis*, 100: 103412, 2025.

Yekutiel Uliel, Alina Ryabtsev, Assaf Hoogi, and Leo Joskowicz. Improving few-shot-segmentation of new structures in volumetric medical images by support set optimization. In *Efficient Medical Artificial Intelligence (EMA4MICCAI 2025)*, volume 16318 of *Lecture Notes in Computer Science*, pp. 41–50, 2026. doi: 10.1007/978-3-032-13961-0_5.

Assefa Seyoum Wahd, Banafshe Felfeliyan, Yuyue Zhou, Shrimanti Ghosh, Adam McArthur, Jiechen Zhang, Jacob L Jaremko, and Abhilash Hareendranathan. Sam2Rad: A segmentation model for medical images with learnable prompts. *Computers in Biology and Medicine*, 187: 109725, 2025.

Jin Wang, Bingfeng Zhang, Jian Pang, Mengyu Liu, Honglong Chen, and Weifeng Liu. Beyond visual cues: Leveraging general semantics as support for few-shot segmentation. *arXiv preprint arXiv:2511.16435*, 2025.

Junwei Wen, Deshui Miao, Guangming Lu, Xin Li, and Wenjie Pei. RCoT-Seg: Reinforced chain-of-thought for video reasoning and segmentation. *arXiv preprint arXiv:2605.07334*, 2026.

Hallee E. Wong, Jose Javier Gonzalez Ortiz, John Guttag, and Adrian V. Dalca. MultiverSeg: Scalable interactive segmentation of biomedical imaging datasets with in-context guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Chenwei Wu, David Restrepo, Zitao Shuai, Zhongming Liu, and Liyue Shen. Efficient in-context medical segmentation with meta-driven visual prompt selection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 255–265. Springer, 2024.

Yang Xing, Jiong Wu, Yuheng Bu, and Kuang Gong. SAM2-SGP: Enhancing SAM2 for medical image segmentation via support-set guided prompting. *arXiv preprint arXiv:2506.19658*, 2025.

Bo Xu, Quanhao Zhu, Rui Lin, Boling Zhu, Chenyuan Wang, Hongfei Lin, Feng Xia, and Chenhua Ji. UltraSAM3: A concept-driven foundation model for universal ultrasound image segmentation. *arXiv preprint arXiv:2607.29200*, 2026.

Chengming Xu, Chen Liu, Yikai Wang, Yuan Yao, and Yanwei Fu. Towards global optimal visual in-context learning prompt selection. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zishan Xu, Yifu Guo, Yuquan Lu, Fengyu Yang, and Junxin Li. VideoSeg-R1: Reasoning video object segmentation via reinforcement learning. *arXiv preprint arXiv:2511.16077*, 2025.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, 2024.

Zhonghao Yan, Muxi Diao, Yuxuan Yang, Ruoyan Jing, Jiayuan Xu, Kaizhou Zhang, Lele Yang, Yanxi Liu, Kongming Liang, and Zhanyu Ma. MedReasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11577–11585, 2026.

Jiewen Yang, Xinpeng Ding, Ziyang Zheng, Xiaowei Xu, and Xiaomeng Li. GraphEcho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11844–11853, 2023.

Zhaoyuan Yin, Jia Zheng, Weixin Luo, Shenhan Qian, Hanling Zhang, and Shenghua Gao. Learning to recommend frame for interactive video object segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with MLLM for dense grounded understanding of images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. arXiv:2501.04001.

Yumin Zhang, Hongliu Li, Yajun Gao, Haoran Duan, Yawen Huang, and Yefeng Zheng. Prototype correlation matching and class-relation reasoning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(12):4041–4054, 2024.

Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. EVF-SAM: Early vision-language fusion for text-prompted segment anything model. *Image and Vision Computing*, 173:106028, 2026. doi: 10.1016/j.imavis.2026.106028.

Chenlan Zhao, Benny Wong, Timothy F. Lundberg, Ahmed M. Elsayed, Abdallah Aljarkas, Hamad A. Aljamaan, Lynn Karam, Qianye Yang, Yipeng Hu, Claire C. Villette, and Shaheer U. Saeed. Active few-shot segmentation by reinforcing data selection. *arXiv preprint arXiv:2607.22371*, 2026. Accepted at EMA4MICCAI 2026, the 2nd MICCAI Workshop on Efficient Medical AI.

Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. SegAgent: Exploring pixel understanding capabilities in MLLMs by imitating human annotator trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3686–3696, 2025.

Yazhou Zhu, Shidong Wang, Tong Xin, and Haofeng Zhang. Few-shot medical image segmentation via a region-enhanced prototypical transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 271–280. Springer, 2023.

Haiyue Zu, Jun Ge, Heting Xiao, Jile Xie, Zhangzhe Zhou, Yifan Meng, Jiayi Ni, Junjie Niu, Linlin Zhang, Li Ni, et al. Rethinking few-shot medical image segmentation by sam2: A training-free framework with augmentative prompting and dynamic matching. *arXiv preprint arXiv:2503.04826*, 2025.

## APPENDIX CONTENTS

- Appendix A: Limitations.
- Appendix B: Datasets, splits and protocol (dataset statistics, unit recovery on TG3K, held-out rules, mirroring, multi-class segmentation).
- Appendix C: USST acquisition and annotation.
- Appendix D: Metrics and evaluation protocol (HD95 ground truth on TG3K, statistics, reference rows, invalid replies).
- Appendix E: Prompts.
- Appendix F: Implementation details (candidate list, both decisions, language models, de-identification, compute).
- Appendix G: Baseline configurations.
- Appendix H: Orientation of the supports on TG3K.
- Appendix I: Anchor decision from raw frames.
- Appendix J: Matcher on TG3K.
- Appendix K: Choice of language model.
- Appendix L: Anatomical reading of the query.
- Appendix M: Reasoning effort of the API model.
- Appendix N: Full results.
- Appendix O: Additional visualizations.

## A LIMITATIONS

- **Level 2 on USST.** Level 2 on USST used 6 of the 18 test patients (evenly spaced), whose contiguous vertebral segments were joined into one sweep per patient; all three support chains ran on these sweeps (Table 2). A joined sweep passes several vertebrae whose shape repeats, which favours several starting frames over a single anchor. Only the chain with DINOv2 supports also ran on the patients' 14 single vertebral segments. With DINOv2 supports on the 14 single vertebral segments, multi-start was also the best policy (all $K$ frames 0.703, LLM anchor 0.671, middle frame 0.658, anchor oracle 0.726). Results of the full system on single vertebral segments are missing.

# B DATASETS, SPLITS AND PROTOCOL

Table 4: Datasets and splits. *Unit*: the grouping used for leave-one-unit-out splits. *Pool*: annotated frames a query may draw supports from (before excluding its own unit). *Queries*: frozen test queries. *Mirror*: mirrored supports allowed (side rule). *L2*: evaluated at Level 2 (anchor selection and propagation); number of test sweeps. *USST: one sweep per patient, joined from its vertebral segments.

| Dataset | Access | Organ (view) | Acquisition | Unit | #Units | Pool | Queries | Mirror | L2 |
|---|---|---|---|---|---|---|---|---|---|
| TG3K | public | thyroid (transverse) | freehand sweep | scan | 21 | 168 | 260 | ✓ | 13 |
| MicroSeg | public | prostate (sagittal) | rotational sweep | patient | 75 | 600 | 355 | – | – |
| CardiacUDA | public | heart (A4C, 4 classes) | cine loop | video | 169 | 150 (Site G) | 100 (Site R) | – | – |
| USST | private | spine (transverse) | freehand sweep | vertebra | 59 | 472 | 250 | – | 6* |

**TG3K.** The frames come from 16 transverse thyroid videos; we used the $256\times256$ copy released in US30K (Lin et al., 2024). The release has no video identifiers, so we recovered 21 continuous scan units from drops in the correlation between consecutive frames (0.997 within a scan, bimodal with a gap between 0.8 and 0.95). Thirteen long scans (at least 100 frames each) served as query units, with 20 evenly spaced queries each. Because the recovered units are not verified patient identities, the pool of a query excluded its own unit and the two neighbouring units.

**MicroSeg.** The release provides one volume per patient (55 training and 20 test patients). We converted the volumes to frames in the orientation of the authors' preprocessing and kept the 2,621 slices with a non-empty expert mask (21–45 per patient); the annotations of the non-expert readers were not used. The slice index is a lateral position (from the midline to the lateral edge of the gland), not base, mid or apex, and its sign cannot be recovered from the data. Seventy-one patients served as query units, with 5 evenly spaced queries each; all 75 contributed to the pool.

**CardiacUDA.** We converted the released volumes to frames. The release annotates most videos on five frames each and a small set of videos on every frame; we used only the sparsely annotated videos of Sites G and R, and not the densely annotated ones. Label values 1–4 denote the left ventricle, left atrium, right atrium and right ventricle. The pool holds 150 frames of 50 Site G videos (the middle frame of each run of annotated frames), and the 100 test queries come from 100 Site R videos, so the split is cross-site as well as cross-patient.

**USST.** USST contains freehand transverse sweeps along the thoracic spine of 21 patients. Each sweep was divided into vertebral units (T5–T11), giving 59 units and 2,980 annotated frames; the label covers the paired vertebral-arch structures on either side of the midline. Every unit was a query unit with 5 evenly spaced queries, and 50 units from 18 patients form the test set (250 queries). Acquisition and annotation are described in Appendix C. At Level 2, 6 of the 18 test patients were used: their 14 single vertebral segments with DINOv2 supports, and, for the full system, one sweep per patient joined from its contiguous vertebral segments.

**Held-out rules.** A query never drew supports from its own unit. The query list was frozen, and every prompt fixed, before the reported runs. The supervised reranker was trained with cross-validation by unit: the model applied to a test unit never saw that unit (on TG3K, nor its neighbours).

**Mirroring on TG3K.** Scans of the left and right thyroid lobe are mirror images of each other, and a support from the opposite side transfers markedly worse. Each pool frame was therefore also available mirrored. A support was mirrored when its gland side, computed from its own mask, was opposite to the query side that the model reported (the *side rule*); mirror flags produced by the model were ignored. No other dataset used mirroring.

**Multi-class segmentation on CardiacUDA.** Each chamber was segmented as a separate binary task from the same supports, and every metric was averaged over the four chambers.

## C USST: Acquisition and Annotation

USST is a private dataset of freehand transverse ultrasound sweeps along the thoracic spine, collected under institutional review board approval and de-identified before use. During each sweep the scanner recorded the position of the probe in millimetres, which allowed the frames to be placed along the spine. The annotation proceeded in five steps.

1. **Coarse segmentation.** Each whole-spine sweep was divided into 19 segments, and each segment received a quality label.
2. **Initial annotation.** On the selected segments, an annotator drew the mask on keyframes of their choice, and SAM2 propagated these masks to every frame of the segment.
3. **Re-segmentation with 3D information.** The masks were placed in 3D using the recorded probe positions and the pixel intensities, and the annotator moved the boundaries between vertebrae in this view (Figure 4). The masks of step 2 followed the new boundaries.
4. **Smoothing.** Every mask was converted into a smooth polygon (about 38 vertices for large shapes and 22 for small ones); the hand-drawn keyframes were kept as drawn.
5. **Frame-by-frame correction.** The annotator corrected the polygons frame by frame, and the final masks were exported.

The final segmentation was therefore made by a human, assisted by 3D information along the spine; the propagated masks served only as proposals that the annotator corrected. The result is 59 vertebral units (T5–T11) from 21 patients with 2,980 annotated frames.

This procedure was one of the sources of Level 2. The annotator chose which frames to draw, and SAM2 propagated them through the segment; Level 2 gives this choice of starting frame to the MLLM.

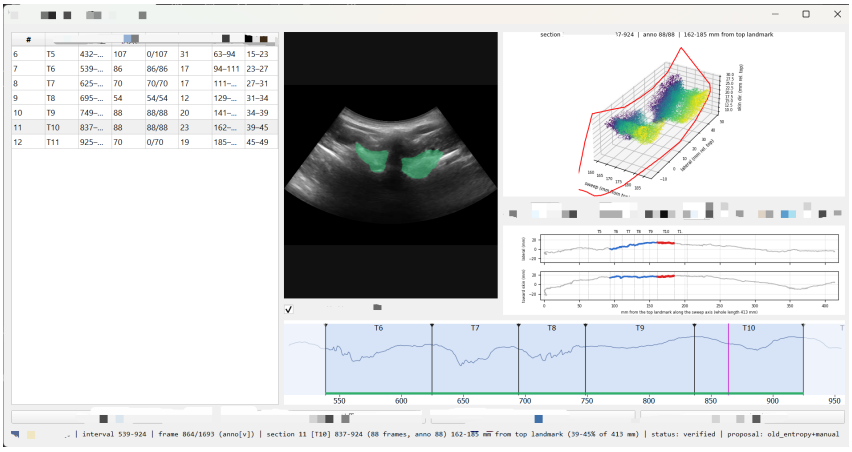



Figure 4: Annotation interface of step 3. Left: the vertebral units of one sweep. Centre: a frame with its mask. Top right: the masks of the unit placed in 3D along the sweep using the recorded probe positions. Middle right: lateral and depth position of the masks along the sweep, with the vertebral boundaries. Bottom: the intensity profile along the sweep with the boundaries that the annotator adjusts.

## D Metrics and Evaluation Protocol

**Segmentation metrics.** Dice, IoU and HD95 were computed by one shared implementation at the native resolution of the query. UniverSeg and MultiverSeg run at $128\times128$; their probabilities were upsampled bilinearly and thresholded at 0.5. HD95 (in pixels) is the larger of the two directed 95th-percentile distances between the surfaces of the prediction and the ground truth. An empty prediction was assigned the image diagonal as its HD95, so that HD95 and Dice are averaged over the same queries. On CardiacUDA every metric was computed per chamber and averaged over the four chambers.

**Ground truth for HD95 on TG3K.** The TG3K masks we used contain many grey levels from interpolation, and binarising them leaves small specks (median 8 px, 0.97% of the mask area) whose boundaries make up more than 5% of the ground-truth boundary. Under the raw ground truth, HD95 barely separated good from bad predictions (oracle 84.6 vs. random 89.2 px; rank correlation with Dice 0.67). For HD95 on TG3K we therefore kept only the 8-connected components of the ground truth whose area is at least 2% of the largest component (rank correlation 0.96), and applied this to every method alike. Dice and IoU used the raw ground truth, which changed them by at most 0.004. The masks of the other datasets were not modified.

**Statistics.** We report the mean over test queries with 95% confidence intervals from a cluster bootstrap over units (4,000 resamples), since queries from one unit are correlated. Methods were compared on the same queries with the paired Wilcoxon signed-rank test. To compare datasets of different difficulty, we also report the *recovered headroom* $(\mathrm{Dice}_{\mathrm{method}} - \mathrm{Dice}_{\mathrm{random}})/(\mathrm{Dice}_{\mathrm{oracle}} - \mathrm{Dice}_{\mathrm{random}})$.

**Reference rows.** These rows use the query's ground truth and are not deployable. The *oracle* greedily picks the Dice-maximising support set from the whole pool; the *union-oracle* does the same within the candidate list shown to the model; *same-video* supports are $n$ annotated frames from the query's own unit. At Level 2, the unit of evaluation is a sweep and the metric is the mean Dice over all its frames after propagation. Its references are the best of the $K$ candidate anchors (anchor oracle) and propagation from the ground-truth mask of the best anchor.

**Invalid replies and failed calls.** For language-model runs we report the rate of invalid replies. An invalid reply got one retry; a second failure was counted and scored with the expected Dice of a uniformly random set from the candidate list. Any failed call (authentication, network, quota or server error) aborted the run and was never replaced by a random choice.

## E PROMPTS

We give the TG3K prompts verbatim; placeholders in braces are filled per query ($k$: number of candidates, $n$=4, last = $k$+1). The prompts for the other datasets have the same structure, with the anatomy of the target and without mirroring; they are released with the code. Images are passed in the order stated in each prompt.

**Exemplar mode (LucidSeg$_{\mathrm{E}}$), query shown.**

```
You are an expert in thyroid ultrasound.
Image 1 is the query: a transverse B-mode ultrasound frame of the neck whose thyroid gland is
      NOT segmented.
Images 2 to {last} are {k} annotated examples from other scans (Example 1 to Example {k}; the
      number is printed in the lower-right corner of each image). In each example the thyroid
      gland is outlined in green.

A segmentation model will segment the query's thyroid gland using {n} examples that you
     choose (each example image together with its gland outline). It transfers an outline
     well when the example shows the gland at a similar position, size and shape in the image
     , with the same neighbouring structures on the same sides. It cannot rotate or move an
     example, but it can mirror an example left-right. Scans of the left and the right lobe
     are mirror images of each other: the common carotid artery (a round dark circle) lies on
      the outer side of the lobe, the trachea (a curved bright arc with a dark acoustic
     shadow below it) on the inner side. Mirror an example when its carotid artery and
     trachea lie on the opposite sides from those in the query.

Steps:
1. In the query, decide on which side of the image the gland lies and on which side of the
     gland the carotid artery lies.
2. Choose {n} different examples that together will segment the query best, and for each say
     whether to mirror it. Prefer examples whose gland, after mirroring if needed, matches
     the query in position, size and shape; avoid examples whose outline would clearly not
     fit.

Return JSON only:
{"query_gland_side": "left" or "right" or "center", "query_carotid_side": "left" or "right"
     or "none", "choices": [{"example": <number 1 to {k}>, "flip": true or false}, ...
     exactly {n} items with different numbers], "reason": "<one or two sentences>"}
```

**Exemplar mode, query withheld.**

```
You are an expert in thyroid ultrasound.
Images 1 to {k} are annotated examples of transverse B-mode ultrasound frames of the neck
     from different scans (Example 1 to Example {k}; the number is printed in the lower-right
      corner of each image). In each example the thyroid gland is outlined in green.

A segmentation model will use {n} examples that you choose (each example image together with
     its gland outline) as references for segmenting the thyroid gland in new frames that you
      cannot see.
Choose the {n} different examples that are the clearest and most representative references.

Return JSON only:
{"choices": [{"example": <number 1 to {k}>}, ... exactly {n} items with different numbers], "
     reason": "<one sentence>"}
```

**Transfer mode (LucidSeg$_{\mathrm{T}}$).**

```
You are an expert in thyroid ultrasound.
Image 1 is the query: a transverse B-mode ultrasound frame of the neck whose thyroid gland is
      NOT segmented.
Images 2 to {last} show the same query frame again, once per annotated example from another
     scan (Example 1 to Example {k}; the number is printed in the lower-right corner). In
     each of them, the thyroid outline of that example has been transferred onto the query
     after matching its position and scale automatically: the GREEN outline is the example as
      it is, the MAGENTA outline is the example mirrored left-right. The automatic matching
     can be wrong, so an outline may be misplaced.

A segmentation model will segment the query's thyroid gland using {n} examples that you
     choose. It works best with examples whose outline, in the better of the two orientations
     , lies on the query's thyroid gland and follows its shape and size.

Steps:
1. Look at the query and find its thyroid gland.
2. Choose the {n} different examples whose green or magenta outline fits the query's thyroid
     gland best, and for each say which orientation fits better ("flip": true for the magenta
      outline).

Return JSON only:
{"choices": [{"example": <number 1 to {k}>, "flip": true or false}, ... exactly {n} items
     with different numbers], "reason": "<one or two sentences>"}
```

**Level 2 anchor decision.** With {domain}, {sweep} and {target} set per dataset (e.g., "thyroid ultrasound", "transverse ultrasound sweep of the neck", "the thyroid gland"):

```
You are an expert in {domain}.
The {k} images are frames, in time order (Frame 1 to Frame {k}; the number is printed in the
     lower-right corner), from one {sweep}. In each frame an automatic segmentation of {
     target} is outlined in green; it was transferred from annotated examples of other
     patients and may be wrong.
We will keep the segmentation of ONE of these frames and track it through the whole sweep in
     both directions, so an error in the chosen frame spreads to every frame.
Choose the ONE frame whose green outline is the most correct: it covers {target} completely,
     follows its borders closely, has a typical shape and includes no other structure. Frames
      near the ends of a sweep often show {target} small or incomplete.
Return JSON only: {"choice": <number 1 to {k}>, "reason": "<one sentence>"}
```

## F Implementation Details

**Candidate list.** For each query, the candidates were the union of the 6 best allowed pool frames by DINOv2 similarity, the 6 best by template matching and 6 random allowed frames (14–18 unique frames, since the lists overlap), shuffled once per query with a fixed seed. Each candidate was rendered at 512 px with its annotation outline in green and its number printed in the lower-right corner; the query was rendered without an outline. The main setting gave no text descriptions (they are ablated in Section 6). We logged the provenance of each pick (DINOv2 list, template-matching list or random part) to test whether the model merely followed a visual criterion.

**Support decision (Level 1).** The model received the query and all candidates in one call and returned a JSON object with $n{=}4$ distinct candidate indices. We used two ways of presenting the candidates. *Exemplar mode* (LucidSeg$_{\mathrm{E}}$). Each candidate was shown as its own image with its annotation outline. The model also returned a short anatomical reading of the query (the gland side

on TG3K, the lateral position on MicroSeg). On TG3K the chosen supports were mirrored by the side rule (Appendix B). In the *no-query* control of this mode the query image was withheld and the model chose the $n$ most representative examples. *Transfer mode* (LucidSeg$_{\mathrm{T}}$). Each candidate's annotation was aligned to the query by the template matching of Appendix G (translation and scale only) and drawn on a copy of the query, one tile per candidate; on CardiacUDA the four chambers were drawn in four colours. The model chose the $n$ candidates whose transferred outline fits the query's structure best, knowing that the automatic alignment can be wrong. On TG3K each tile showed both orientations (the candidate as is and mirrored, in two colours), and the model's choice of colour set the orientation. Only the choice of supports is passed on: the executor receives the original support images and masks, not the aligned outlines.

**Anchor decision (Level 2).** We used $K{=}8$ candidate anchors. Every candidate was segmented by MultiverSeg from $n{=}4$ supports and drawn at 512 px with its predicted outline in green and its number in the lower-right corner. SAM2.1-Large propagated the anchor mask forwards and backwards through the sweep. To isolate the anchor decision, we first fixed the supports to DINOv2 and let Gemma-4-31B choose the anchor. In the *full system*, one model made both decisions; we ran it with Gemma-4-31B and with GPT-5.5 without reasoning.

**Language models.** Because clinical ultrasound is subject to privacy constraints, the main setting used locally deployed open-weight models, Qwen3.5-9B and Gemma-4-31B, served with vLLM at temperature 0, with JSON-constrained decoding and thinking disabled. The closed models GPT-5.5 and Claude Sonnet 5 were evaluated on the same test queries as the local models. Claude Sonnet 5 kept its default reasoning setting. GPT-5.5 ran without reasoning (effort "none"); Appendix M compares it with its default reasoning.

**De-identification.** Before any image was sent to a closed model, everything outside the imaged sector was masked, which removes burned-in text such as scanner labels, and images were re-encoded in memory. Prompts contained no identifiers, only example numbers. USST images were sent only to OpenAI and Anthropic, for research use approved for this dataset.

**Executors and features.** UniverSeg and MultiverSeg ran with their public weights, without fine-tuning. DINOv2 ViT-B/14 provided the features for the visual selectors and the candidate list.

**Compute.** All runs used NVIDIA RTX A6000 GPUs (48 GB). One decision call took about 5–6 s with Qwen3.5-9B and 12–16 s with Gemma-4-31B.

## G Baseline Configurations

Every baseline ran in its own environment with its official weights and default settings; at most one adjustment was made per method and is documented below. All methods saw the same queries and pool and were scored by the same implementation.

**Visual in-context segmenters.** UniverSeg (Butoi et al., 2023) and MultiverSeg (Wong et al., 2025) took the $n{=}4$ supports as their context set at $128{\times}128$; their probabilities were upsampled to the native resolution before thresholding. Matcher is described in Appendix J.

**Visual selection criteria.** All criteria chose among the pool frames allowed for the query (on TG3K, each frame as is or mirrored). *DINOv2 similarity* ranked frames by the cosine similarity between the CLS embeddings (ViT-B/14) of the query and the frame. *Template matching* cropped the annotated structure of each pool frame with a margin of 8% of its size, slid the crop over the query at several scales after downsampling and smoothing both images, and ranked frames by the highest zero-mean normalized cross-correlation. The *reranker* was a gradient-boosted regression model (200 iterations, learning rate 0.05, at most 15 leaves) that predicted the single-support Dice of a candidate from these two scores and, on TG3K, those of its mirrored copy; the $n$ candidates with the highest prediction were chosen. It was trained with cross-validation by unit: the model applied to a test unit was trained only on other units (on TG3K, excluding its neighbours).

**Text-driven models.** GroundingDINO-US-SAM (Rasaee et al., 2025) and UltraSAM3 (Xu et al., 2026) were prompted with the class names of their own training data. For UltraSAM3 this means its own mapping from label value to name; on CardiacUDA its training scripts name the four labels in an order that does not match the anatomy, and we kept that mapping because it is the model's interface. Each chamber was prompted as a separate binary task. The released weights of GroundingDINO-US-SAM were trained on TG3K and MicroSeg, and those of UltraSAM3 on TG3K (with a random split of frames), MicroSeg and CardiacUDA, so our test frames may be in their training data; on these datasets their results are in-domain references, not zero-shot results. USST is unseen by both, and CardiacUDA by GroundingDINO-US-SAM.

**Agent-based model.** VisionReasoner (Liu et al., 2026b) ran zero-shot with its official weights; the query text named the target structure, and its box and point outputs prompted a frozen SAM2 (Hiera-L) as in its official pipeline (Qwen2.5-VL-7B backbone, greedy decoding). On CardiacUDA each text-driven or agent model was evaluated in its training orientation: UltraSAM3, which was fine-tuned on CardiacUDA as stored, on the frames as stored; GroundingDINO-US-SAM and VisionReasoner on the standard apical four-chamber display (frames mirrored for inference and predictions mirrored back). The two orientations gave nearly the same Dice for the latter two.

## H Orientation of the Supports on TG3K

On TG3K, scans of the left and right lobe are mirror images, so each chosen support must also be given an orientation. We compared the side rule of Section 3.2, in which the model reports only the side of the query, with letting the model decide the orientation of each support itself (Table 5). All rows use the same exemplar-mode prompt and differ only in how the orientation is set.

Table 5: Orientation of the chosen supports on TG3K (LucidSeg$_{\mathrm{E}}$; Dice UniverSeg / MultiverSeg; each model's four supports are fixed and only the orientation changes).

| Orientation decided by | Qwen3.5-9B | Gemma-4-31B | GPT-5.5 (reasoning) | GPT-5.5 | Sonnet 5 |
|---|---|---|---|---|---|
| No mirroring | .457 / .592 | .460 / .588 | .466 / .594 | .456 / .572 | .423 / .562 |
| Model's own mirror flags | .446 / .591 | .473 / .593 | .547 / .659 | .470 / .589 | .417 / .538 |
| Side rule (ours) | .465 / .585 | .484 / .611 | .517 / .642 | .498 / .618 | .417 / .551 |
| Side rule, ground-truth side | .519 / .653 | .514 / .649 | .530 / .662 | .511 / .638 | .483 / .626 |

The side rule helped where the model read the side of the query reliably: Gemma-4-31B and GPT-5.5 without reasoning gained 0.018 and 0.029 with MultiverSeg (0.011 and 0.028 with UniverSeg) over their own mirror flags. It helped little for Qwen3.5-9B ($+0.019$ with UniverSeg, $-0.006$ with MultiverSeg) and Sonnet 5 (0.000 and $+0.013$), whose side readings were wrong in 37% and 53% of the clear queries (Table 9). With its default reasoning, GPT-5.5 set the orientation better by itself than through the rule (0.659 vs. 0.642). Every model stayed 0.01–0.08 below the rule with the ground-truth side, so the remaining error lies in reading the side, not in the rule. The main tables use the side rule for every model.

## I Anchor Decision from Raw Frames

In the main text, each candidate anchor is shown to the model with the executor's predicted outline. Here we compare this with a variant that shows the eight candidate frames without any outline and asks for the best starting frame from their appearance alone. Table 6 compares the two on the same sweeps, with the non-LLM anchors for reference.

From the raw frames alone, the MLLM never chose a significantly better anchor than the middle frame (at most $+0.005$, GPT-5.5 on TG3K), and in the Gemma-4-31B chain on TG3K it was below the executor's confidence ($-0.047$, $p{=}0.027$). With the predicted outlines, its choice improved by 0.03–0.06 in every chain on TG3K and by up to 0.11 on USST. The anchor decision thus relies on the result of the support decision.

Table 6: Anchor decision from raw frames vs. from frames with the predicted outline (Dice over all frames of a sweep after propagation; executor MultiverSeg; the anchor is chosen by the model of the support chain, by Gemma-4-31B for DINOv2 supports).

| Dataset | Supports by | $N$ | LLM, raw frames | LLM, with prediction | Middle | Confidence | Random (expected) | Anchor oracle |
|---|---|---|---|---|---|---|---|---|
| TG3K | DINOv2 | 13 | .718 | .780 | .745 | .737 | .566 | .828 |
| | Gemma | 13 | .755 | .781 | .772 | .802 | .639 | .825 |
| | GPT-5.5 | 13 | .740 | .785 | .735 | .754 | .612 | .833 |
| USST | DINOv2 | 6 | .568 | .678 | .623 | .589 | .576 | .706 |
| | Gemma | 6 | .639 | .665 | .640 | .630 | .609 | .706 |
| | GPT-5.5 | 6 | .581 | .632 | .633 | .622 | .564 | .695 |

## J MATCHER ON TG3K

Matcher (Liu et al., 2024) matches DINOv2-L patch features between the supports and the query and prompts SAM-H with the matched points. We ran the official code and weights, unmodified, in their own environment, under the same protocol as the main experiments (260 TG3K test queries, the same pool and exclusions). Images were resized to $518\times518$ as in its few-shot loader and predictions resized back to $256\times256$; Dice and IoU used the raw ground truth and HD95 the cleaned one (Appendix D). We evaluated $n{=}1$ and $n{=}4$ supports, the latter through Matcher's own $n$-shot mode, chosen at random, by DINOv2 similarity, by the reranker, or by LucidSeg (Gemma-4-31B with the side rule).

**Settings.** We report the better of the two official presets (the FSS preset) and one documented ultrasound adaptation, which uses only Matcher's own parameters and the support annotations: no merging of SAM masks, and an area prior that discards SAM proposals larger than twice the largest support mask. The adaptation was fixed before the test runs. As a diagnostic upper reference, we also used the query itself as the support (*self-support*); this is not a baseline.

Table 7: Matcher on TG3K (260 test queries). Dice with 95% cluster-bootstrap CI, IoU, and HD95 (px, cleaned ground truth).

| Setting | Supports | Dice [95% CI] | IoU | HD95 |
|---|---|---|---|---|
| Official (FSS preset) | $n{=}1$, random | 0.259 [0.218, 0.304] | 0.165 | 123.5 |
| | $n{=}1$, DINOv2 | 0.303 [0.250, 0.357] | 0.202 | 116.5 |
| | $n{=}4$, random | 0.278 [0.232, 0.328] | 0.179 | 121.4 |
| | $n{=}4$, DINOv2 | 0.326 [0.265, 0.391] | 0.222 | 110.6 |
| | self-support | 0.510 [0.430, 0.594] | 0.393 | 76.8 |
| Ultrasound-adapted | $n{=}1$, random | 0.322 [0.270, 0.379] | 0.230 | 108.6 |
| | $n{=}1$, DINOv2 | 0.395 [0.325, 0.467] | 0.295 | 96.2 |
| | $n{=}4$, random | 0.355 [0.297, 0.421] | 0.250 | 102.8 |
| | $n{=}4$, DINOv2 | 0.397 [0.320, 0.479] | 0.292 | 94.7 |
| | $n{=}4$, reranker | 0.347 [0.289, 0.412] | 0.244 | 104.2 |
| | $n{=}4$, LucidSeg (Gemma-4-31B) | 0.385 [0.321, 0.453] | 0.280 | 96.5 |
| | self-support | 0.594 [0.527, 0.666] | 0.473 | 64.1 |

**Findings.** Matcher is a weak executor on ultrasound: even with the query as its own support it reaches only 0.510 (official) and 0.594 (adapted) Dice. The bottleneck is the generation and merging of SAM masks. Under the official presets, 85–98% of the predictions are more than twice the size of the ground truth, often covering the whole imaging sector. The adaptation improves every configuration by 0.06–0.09 Dice ($p < 10^{-12}$). Because the executor compresses the differences between support sets, Matcher supports only a limited conclusion. With the adapted setting and $n{=}4$, the supports chosen by LucidSeg beat random supports ($+0.030$, better on 62% of queries, $p = 4.1\times10^{-4}$) and the reranker ($+0.038$, 65%, $p = 1.2\times10^{-6}$). The reranker, trained on UniverSeg Dice, does not transfer to Matcher and is no better than random. LucidSeg and DINOv2 similarity are tied ($-0.011$, $p = 0.21$), and DINOv2 beats random in all four settings ($+0.04$ to $+0.07$, $p \leq 0.007$). Whether LucidSeg or DINOv2 similarity chooses better supports is therefore decided on UniverSeg and MultiverSeg (Section 5.1), not on Matcher.

## K LANGUAGE MODELS

Table 8 compares the four MLLMs in both modes. Gemma-4-31B and GPT-5.5 were the strongest and the most consistent across datasets. The two weaker models failed in ways that match their anatomical reading: Sonnet 5 read the side of the thyroid query correctly in fewer than half of the clear cases, and its exemplar mode fell below the best visual policy on TG3K; Qwen3.5-9B lost 0.10 on TG3K in the transfer mode relative to its exemplar mode. No model returned an invalid reply after one retry.

Table 8: **Choice of language model** (Dice with UniverSeg; E: exemplar mode, T: transfer mode; on TG3K, E uses the side rule). The local models ran with thinking disabled, GPT-5.5 without reasoning (compared with its default reasoning in Appendix M), and Claude Sonnet 5 with its default setting. No reply was invalid after one retry for any model.

| | | TG3K | | MicroSeg | | CardiacUDA | | USST | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Access | E | T | E | T | E | T | E | T |
| Qwen3.5-9B | open | .465 | .370 | .572 | .662 | .675 | .698 | .560 | .609 |
| Gemma-4-31B | open | .484 | .508 | .694 | .674 | .665 | .703 | .576 | .604 |
| GPT-5.5 | API | .498 | .497 | .698 | .663 | .687 | .703 | .593 | .593 |
| Claude Sonnet 5 | API | .417 | .459 | .600 | .656 | .700 | .688 | .562 | .585 |
| Best visual-based policy | – | .472 | | .718 | | .688 | | .616 | |

## L ANATOMICAL READING OF THE QUERY

On TG3K, the exemplar mode asks the model for the side of the thyroid gland in the query, and the side rule turns this reading into the orientation of the supports (Section 3.2). Table 9 reports how often the side was read correctly and how much Dice the remaining errors cost: replacing the reported side by the ground-truth side raised Dice by 0.01–0.08.

Table 9: **Anatomical reading of the query** (TG3K, LucidSeg$_{\mathrm{E}}$). *Side acc.*: the gland side reported by the model vs. the side of the query's mask centroid, on the 156 queries with a clear side. Dice UniverSeg / MultiverSeg with the model's four supports fixed, mirrored by the side rule with the reported or the ground-truth side.

| | Qwen3.5-9B | Gemma-4-31B | GPT-5.5 (reasoning) | GPT-5.5 | Sonnet 5 |
|---|---|---|---|---|---|
| Side accuracy (%) | 63 | 68 | 76 | 77 | 47 |
| Dice, reported side | .465 / .585 | .484 / .611 | .517 / .642 | .498 / .618 | .417 / .551 |
| Dice, ground-truth side | .519 / .653 | .514 / .649 | .530 / .662 | .511 / .638 | .483 / .626 |

## M REASONING EFFORT OF THE API MODEL

We tested whether long reasoning helps the support decision. GPT-5.5 received the same prompts, test queries and candidate lists; only the OpenAI reasoning effort changed, from the model's default to *none*. With the default setting a call used about 4,900 input and 2,200 output tokens, of which about 2,100 were reasoning tokens. Without reasoning the output was about 100 tokens, so the output cost fell by more than an order of magnitude; over all our GPT-5.5 runs (about 13.6M input and 5.7M output tokens), reasoning accounted for about 93% of the output tokens. No call failed in either setting.

Most differences between the two settings were within $\pm 0.02$ Dice; in the transfer mode they were not significant ($p \geq 0.61$, paired Wilcoxon over queries). The one large difference was the exemplar mode on TG3K with the query shown ($-0.077$ / $-0.070$ without reasoning, $p = 3{\times}10^{-8}$ / $3{\times}10^{-5}$). This is the only setting in which the model also decides the orientation of the supports. Without reasoning it mirrored 11% of the chosen supports instead of 30%, and its orientations agreed with

Table 10: GPT-5.5 with and without reasoning (Dice, UniverSeg / MultiverSeg; 95% CIs in Table 11). E: exemplar mode with the query shown; E, no query: the query withheld; T: transfer mode. On TG3K, E uses the model's own mirror flags. –: not run for cost reasons (on USST the run with reasoning was stopped after 25 of 250 calls). The main text uses the setting without reasoning.

| | E | | E, no query | | T | |
|---|---|---|---|---|---|---|
| Dataset | none | default | none | default | none | default |
| TG3K | 0.470 / 0.589 | 0.547 / 0.659 | 0.468 / 0.601 | 0.462 / 0.593 | 0.497 / 0.633 | 0.505 / 0.639 |
| MicroSeg | 0.698 / 0.754 | 0.699 / 0.772 | 0.657 / 0.732 | 0.666 / 0.739 | 0.663 / 0.738 | – |
| CardiacUDA | 0.687 / 0.776 | 0.694 / 0.785 | 0.670 / 0.766 | 0.671 / 0.762 | 0.703 / 0.787 | 0.711 / 0.788 |
| USST | 0.593 / 0.666 | 0.615 / 0.676 | 0.572 / 0.673 | 0.586 / 0.682 | 0.593 / 0.654 | – |

the better one by single-support Dice in 65% of cases, close to never mirroring (62%). The side of the query it reported was as accurate as with reasoning (77% vs. 76%). Replacing its mirror flags by the side rule recovered about 0.03 (0.498 / 0.618), which is still below the default setting with its own flips (0.547 / 0.659); applying the side rule to the default setting lowered it to 0.517 / 0.642. In the transfer mode the mirrored outline is drawn for the model, and without the query there is nothing to mirror; in both cases the advantage of reasoning disappeared. Reasoning therefore did not measurably change which supports the model chose, and mattered only for the geometric judgment of whether to mirror a support. For this reason, and at a fraction of the cost, all GPT-5.5 results in the main text use the setting without reasoning.

## N Full Results

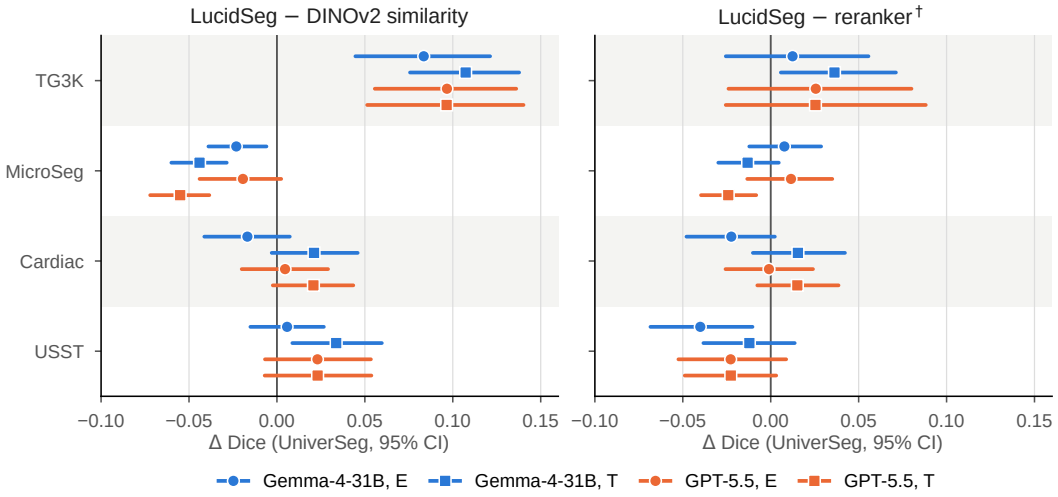



Figure 5: Paired Dice differences of LucidSeg against DINOv2 similarity (left) and against the reranker trained on the target dataset (right), with UniverSeg; mean over test queries and 95% unit-bootstrap confidence interval. Circles: exemplar mode (E); squares: transfer mode (T); blue: Gemma-4-31B; orange: GPT-5.5.

Table 11: 95% unit-bootstrap confidence intervals for Table 10 (GPT-5.5, UniverSeg and MultiverSeg), with reasoning ; without reasoning. –: not run.

| Dataset | Executor | E | E, no query | T |
|---|---|---|---|---|
| TG3K | UniverSeg | [0.487, 0.597] ; [0.406, 0.528] | [0.392, 0.524] ; [0.412, 0.523] | [0.447, 0.554] ; [0.449, 0.537] |
| | MultiverSeg | [0.600, 0.709] ; [0.522, 0.648] | [0.520, 0.657] ; [0.538, 0.657] | [0.589, 0.684] ; [0.575, 0.679] |
| MicroSeg | UniverSeg | [0.680, 0.717] ; [0.672, 0.723] | [0.648, 0.684] ; [0.641, 0.672] | – ; [0.644, 0.680] |
| | MultiverSeg | [0.752, 0.792] ; [0.728, 0.780] | [0.719, 0.760] ; [0.712, 0.752] | – ; [0.717, 0.759] |
| CardiacUDA | UniverSeg | [0.664, 0.723] ; [0.655, 0.717] | [0.636, 0.706] ; [0.636, 0.703] | [0.679, 0.742] ; [0.672, 0.732] |
| | MultiverSeg | [0.760, 0.807] ; [0.750, 0.799] | [0.732, 0.788] ; [0.738, 0.791] | [0.764, 0.810] ; [0.764, 0.809] |
| USST | UniverSeg | [0.582, 0.648] ; [0.558, 0.627] | [0.554, 0.617] ; [0.535, 0.608] | – ; [0.561, 0.625] |
| | MultiverSeg | [0.642, 0.706] ; [0.633, 0.698] | [0.653, 0.709] ; [0.641, 0.704] | – ; [0.620, 0.686] |

Table 12: IoU and HD95 (pixels at native resolution; on TG3K computed on the cleaned ground truth, Appendix D) with UniverSeg. *A single draw of random supports (the draw shown in Figure 3); Table 1 reports the mean over draws. †Trained on the target dataset. Grey: the model was trained on this dataset.

| | TG3K | | MicroSeg | | CardiacUDA | | USST | |
|---|---|---|---|---|---|---|---|---|
| Method | IoU↑ | HD95↓ | IoU↑ | HD95↓ | IoU↑ | HD95↓ | IoU↑ | HD95↓ |
| *Visual-based* | | | | | | | | |
| Random supports* | .264 | 81.9 | .458 | 138.6 | .497 | 61.9 | .386 | 68.3 |
| DINOv2 similarity | .284 | 76.5 | .601 | 101.2 | .552 | 53.1 | .422 | 57.9 |
| Reranker† | .339 | 70.0 | .571 | 111.9 | .562 | 52.8 | .465 | 46.8 |
| *Text-driven and agent-based* | | | | | | | | |
| GroundingDINO-US-SAM | .684 | 49.2 | .602 | 91.2 | .156 | 215.1 | .064 | 203.7 |
| UltraSAM3 | .769 | 23.4 | .704 | 63.1 | .757 | 25.0 | .018 | 149.9 |
| VisionReasoner | .180 | 95.8 | .395 | 145.0 | .100 | 243.0 | .076 | 170.0 |
| *Ours* | | | | | | | | |
| LucidSeg$_{\mathrm{E}}$, Gemma-4-31B | .357 | 67.7 | .579 | 108.9 | .533 | 58.1 | .429 | 55.8 |
| LucidSeg$_{\mathrm{T}}$, Gemma-4-31B | .378 | 68.0 | .565 | 122.1 | .575 | 46.7 | .455 | 56.5 |
| LucidSeg$_{\mathrm{E}}$, GPT-5.5 | .371 | 65.2 | .585 | 105.2 | .557 | 53.5 | .445 | 53.4 |
| LucidSeg$_{\mathrm{T}}$, GPT-5.5 | .367 | 64.3 | .555 | 125.5 | .575 | 48.9 | .446 | 54.8 |
| *Reference (uses ground truth)* | | | | | | | | |
| Oracle (whole pool) | .774 | 27.6 | .874 | 37.5 | .780 | 22.0 | .788 | 11.8 |

## O Additional Visualizations

All panels show predictions in green and the ground truth as a white dashed outline; examples are chosen by fixed rules, including failure cases. Figures 6–9 show Level 1 on each dataset, Figure 10 shows the input of one decision in both modes, and Figure 11 shows three more sweeps at Level 2.

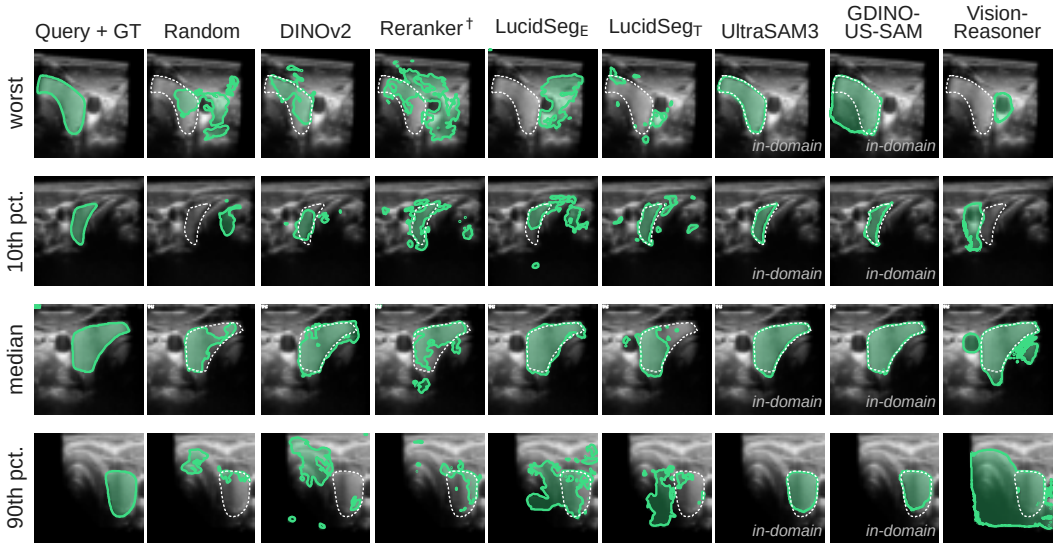



Figure 6: Level 1 on TG3K (UniverSeg; LucidSeg with Gemma-4-31B). Rows: the test queries with the lowest, the 10th-percentile, the median and the 90th-percentile Dice difference between $\text{LucidSeg}_{\text{E}}$ and DINOv2 similarity.

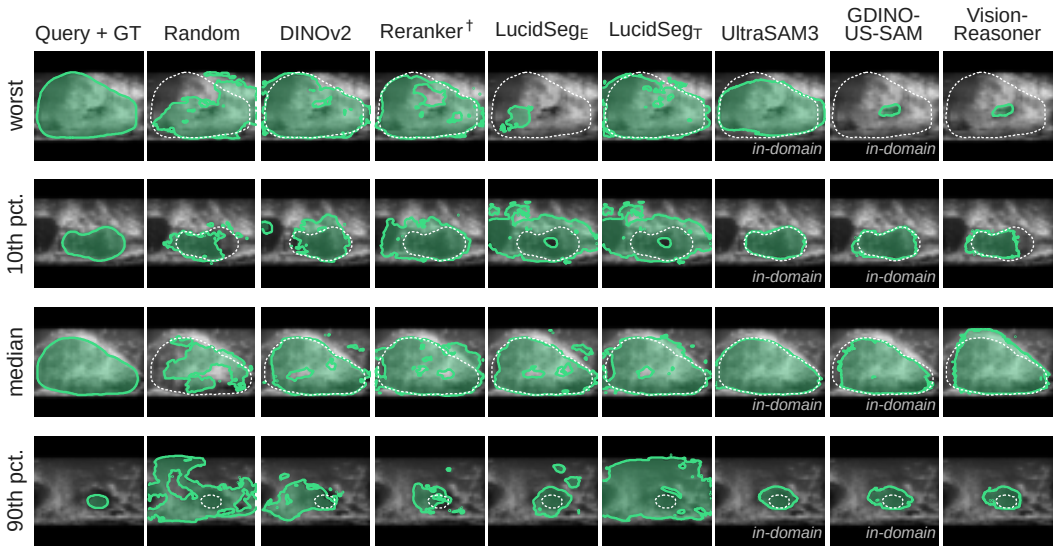



Figure 7: Level 1 on MicroSeg (UniverSeg; LucidSeg with Gemma-4-31B). Rows: the test queries with the lowest, the 10th-percentile, the median and the 90th-percentile Dice difference between $\text{LucidSeg}_{\text{E}}$ and DINOv2 similarity.

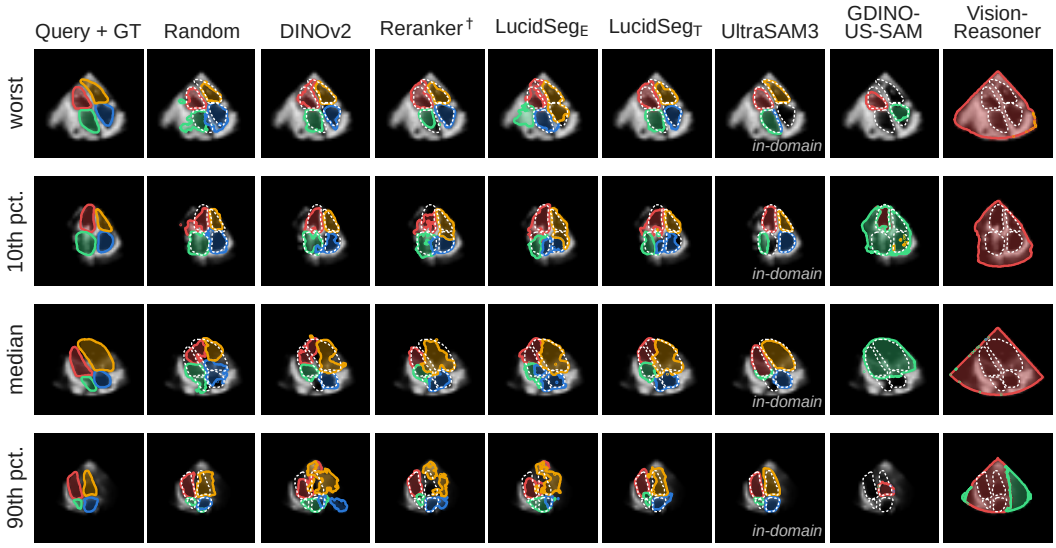



Figure 8: Level 1 on CardiacUDA (UniverSeg; LucidSeg with Gemma-4-31B). Rows: the test queries with the lowest, the 10th-percentile, the median and the 90th-percentile Dice difference between $\text{LucidSeg}_{\text{E}}$ and DINOv2 similarity.

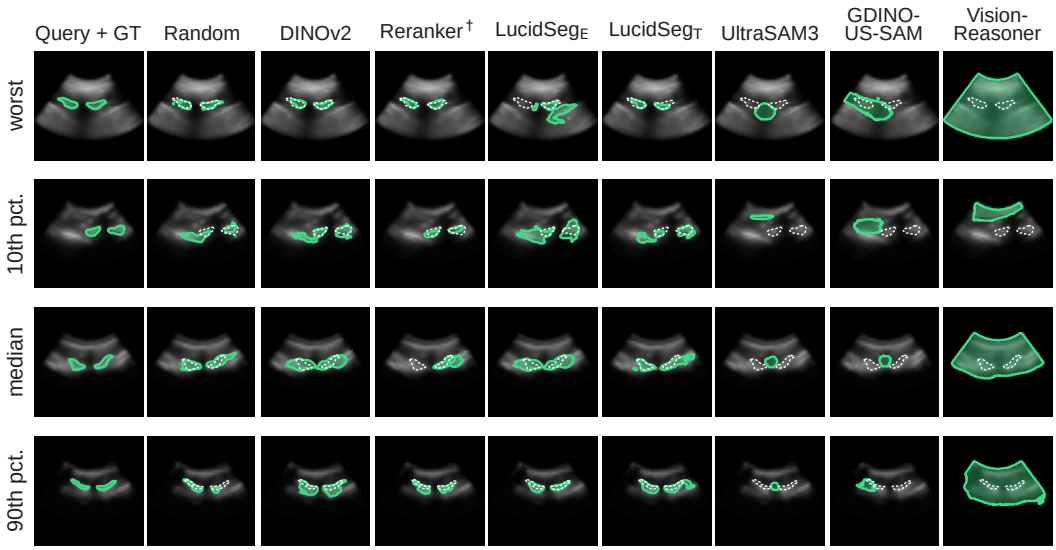



Figure 9: Level 1 on USST (UniverSeg; LucidSeg with Gemma-4-31B). Rows: the test queries with the lowest, the 10th-percentile, the median and the 90th-percentile Dice difference between $\text{LucidSeg}_{\text{E}}$ and DINOv2 similarity.

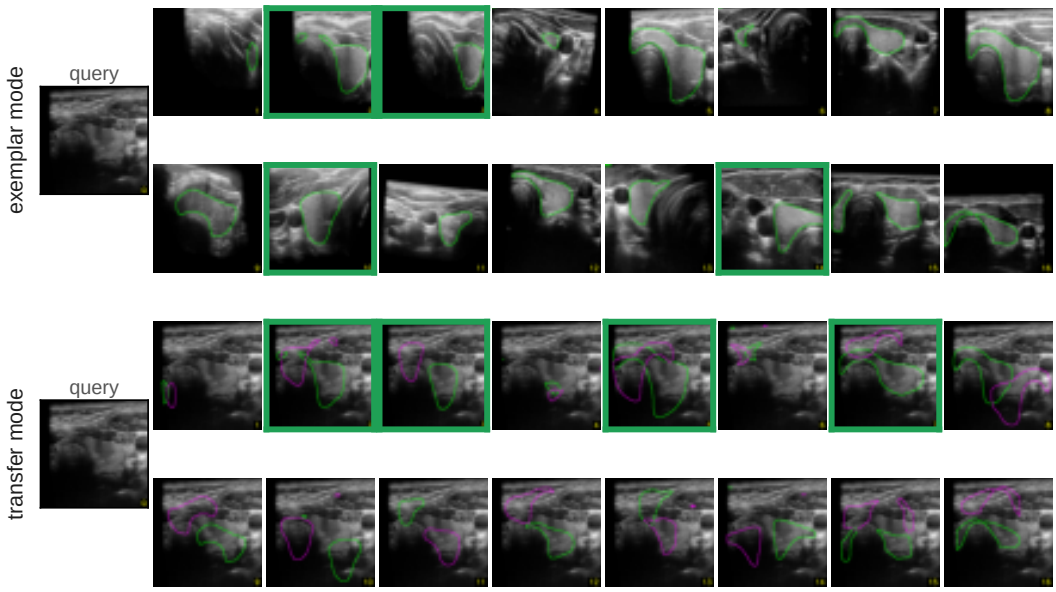
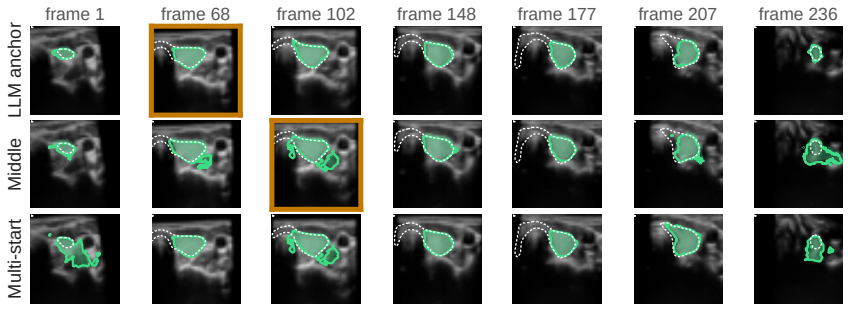
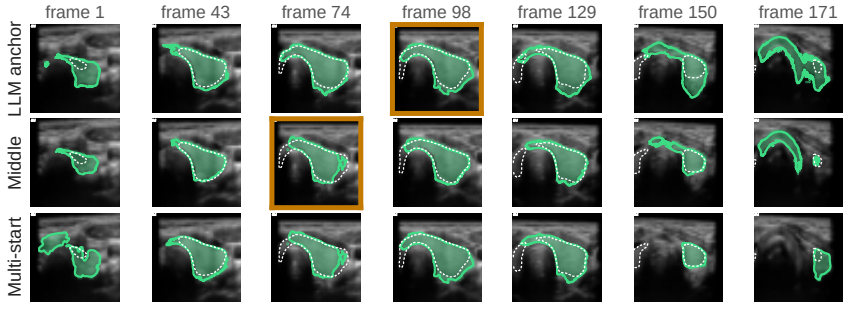


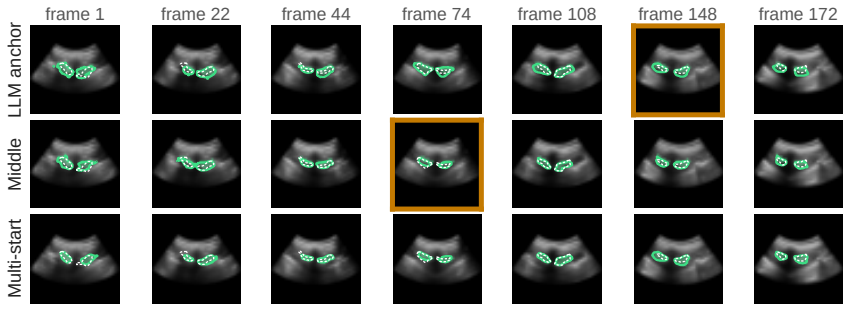

Figure 10: What the MLLM sees for the query of Figure 1 (TG3K): the 16 candidates in the exemplar mode (top) and in the transfer mode (bottom; green: candidate as is, magenta: mirrored). Green frames mark the four candidates chosen by Gemma-4-31B in each mode.




(a)




(b)




(c)

Figure 11: Level 2 on three more test sweeps (full system, Gemma-4-31B, MultiverSeg). Among the sweeps where the LLM did not choose the middle candidate: **(a)** the TG3K sweep with the largest gain and **(b)** the TG3K sweep with the largest loss of the LLM anchor over the middle frame; **(c)** the median USST sweep. The orange border marks the frame from which the row is propagated.