VastFSS: Vast Category Few-Shot Segmentation

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we propose a novel benchmark, dubbed **VastFSS**, aiming to facilitate the development of more generalizable few-shot segmentation (FSS) by covering a vast number of categories across diverse scenarios. Specifically, VastFSS contains 104K images from 2,400 classes, spanning a wide range of application scenarios, from everyday natural scenes to specialized domains, such as biology, agriculture, medical imaging, remote sensing, aerospace science, and other scientific imagery. Compared to existing datasets, our VastFSS provides a broader and more diverse platform for training and evaluating generalizable FSS models, thereby facilitating their real-world applications. In addition, to encourage future research, we present the ***M**ultimodal large language model (MLLM)-**A**ssisted **FSS***, termed **MA-FSS**, a novel framework which leverages the visual understanding capabilities of MLLMs for improving FSS. The key idea is to introduce a set of learnable object tokens to extract target-relevant representations from the support sample via an MLLM and employ them as additional prompts for a segmentation model to predict the query mask. To enhance the representation, we present mask-guided attention within the MLLM to direct the object tokens towards foreground target in the support image, together with a simple adaptive token gate to dynamically adjust their importance, leading to better performance. In extensive experiments on VastFSS, we show that MA-FSS achieves promising performance, outperforming other FSS methods, and generalizes well in a zero-shot manner to existing FSS benchmarks, demonstrating both the effectiveness of MA-FSS and benefits of the broad category and scenario coverage in VastFSS. Our benchmark and code will be released.

1 Introduction

Few-shot segmentation (FSS) aims to segment targets of arbitrary novel categories using a few (*e.g.*, one or five) annotated examples, and is particularly crucial in domains where acquiring dense pixel-level annotations is costly and labor-intensive, such as medical imaging and remote sensing. In the past years, FSS has witnessed considerable progress with many methods proposed (Zou et al., 2025; Xu et al., 2024a; 2025; Karimi & Poullis, 2025; Wen et al., 2025). In this process, FSS benchmarks (Zou et al., 2025; Xu et al., 2024a; 2025; Karimi & Poullis, 2025; Wen et al., 2025; Kavur et al., 2021; Zamir et al., 2019; Zhuang, 2018) have played a crucial role in training and evaluating different methods. Despite this, achieving universally generalizable FSS remains challenging.

The key reason is relatively restricted number of object categories in current FSS benchmarks. The objects in the real world are from *countless* categories. To achieve general few-shot segmentation, FSS models is expected to "SEE" various support-query samples from an extremely large set of object categories during training to acquire the generalization ability. However, widely used standard FSS benchmarks such as PASCAL-5^i (Shaban et al., 2017) and COCO-20^i (Nguyen & Todorovic, 2019) contain only 20 and 80 categories, respectively, as summarized in Fig. 1. And benchmarks developed for specialized applications, including medical imaging and remote sensing like Cardiac MRI (Zhuang, 2018) and iSAID-5^i (Zamir et al., 2019), also typically focus on a limited set of target categories. These restricted vocabularies constrain both the variety of training examples and the scope of generalization assessment. Their limited category vocabularies motivate larger benchmarks such as FSS-1000 (Li et al., 2020) and LVIS-92^i (Liu et al., 2024b), which expand the category space to 1,000 and 920 categories, respectively. Despite these advances, their category vocabularies remain limited relative to the vast diversity of visual concepts encountered in the real world. Universal

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.

FSS requires models to adapt to a wide range of novel categories given only a few annotated examples. Expanding the category vocabulary would therefore enable more comprehensive evaluation of this ability and support the development of more generalizable models, advancing research toward universal FSS. Moreover, category coverage should extend beyond everyday objects to include targets relevant to specialized applications, which introduce additional visual characteristics and segmentation challenges. A unified, large-vocabulary benchmark encompassing both natural-image categories and targets from diverse application contexts would therefore provide richer training opportunities and a more comprehensive testbed for evaluating adaptation to unseen categories. Such a benchmark is an important step toward more generalizable few-shot segmentation.

To address these limitations, we introduce **VastFSS**, a large-vocabulary benchmark comprising 104K images and covering a vast range of 2,400 target categories. These categories encompass everyday objects in natural images as well as targets from specialized applications such as medical imaging, remote sensing, and aerospace science. As shown in Fig. 1, VastFSS substantially expands the category coverage of existing FSS benchmarks. By bringing this broad range of visual concepts into a unified benchmark, VastFSS provides a more comprehensive testbed for developing and evaluating generalizable FSS models, advancing research toward universal few-shot segmentation. To the best of our knowledge, VastFSS is the most comprehensive benchmark for universal few-shot segmentation to date.

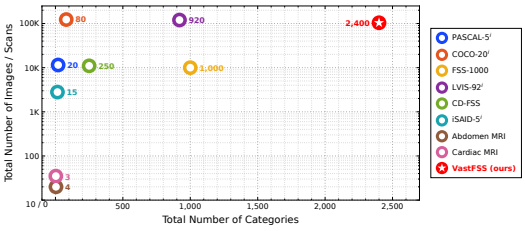


Figure 1: Comparison of VastFSS with existing FSS benchmarks in terms of total categories and dataset size. We can clearly see that VastFSS is *larger* than other datasets by containing 2,400 object categories.

Furthermore, to encourage future research on VastFSS, we further propose ***M**ultimodal large language model (MLLM)-**A**ssisted **FSS*** (**MA-FSS**), a simple yet effective method that uses the visual understanding capabilities of MLLMs (Bai et al., 2025) for few-shot segmentation. Unlike existing MLLM-based segmentation methods (Lai et al., 2024; Zhang et al., 2024c; Yuan et al., 2026; Karimi & Poullis, 2025), which typically rely on explicit linguistic instructions to guide segmentation of the query image, MA-FSS derives the target concept directly from the annotated support sample, thereby adhering to the standard FSS setting in which only support images and corresponding masks are provided. Specifically, MA-FSS introduces a set of learnable tokens into the MLLM to extract target-aware representations from the support image. We further develop a mask-guided attention mechanism that directs theses tokens toward the annotated target region while suppressing irrelevant background information. An adaptive token gate then adjusts the contribution of individual tokens to the target representation. The reweighted tokens serve as target-aware prompts, while a support-guided spatial prior provides complementary localization cues through memory encoding and attention. Together, they guide the mask predictor to segment the corresponding target in the query image. Extensive experiments on VastFSS show that MA-FSS outperforms recent FSS methods, validating its effectiveness of the proposed design and establishing a strong baseline.

Our main contributions are summarized as follows: ♠ We introduce **VastFSS**, a benchmark for universal few-shot segmentation that covers a vast vocabulary of 2,400 categories across 104K images. ♦ We propose MA-FSS, a simple yet effective baseline that leverages MLLM-derived target-aware representations for FSS, offering a strong starting point for future research on VastFSS. ♣ Extensive experiments on VastFSS demonstrate that MA-FSS achieves promising results.

2 Related Work

Few-Shot Segmentation Benchmarks. Benchmarks have been central to the development of few-shot segmentation. PASCAL-5^i (Shaban et al., 2017) and COCO-20^i (Nguyen & Todorovic, 2019) contain 20 and 80 categories, while larger benchmarks such as FSS-1000 (Li et al., 2020) and LVIS-92^i (Liu et al., 2024b) expand the category space to 1,000 and 920 categories, respectively. However, existing benchmarks remain largely focused on natural images, while cross-domain and specialized FSS datasets (Lei et al., 2022; Tong et al., 2024; Sun et al., 2026; Huang et al., 2023; Ouyang et al., 2020; Bi et al., 2025) typically cover only a limited number of target categories. Thus, a large-

vocabulary benchmark spanning both natural and specialized domains remains underexplored. To fill this gap, we introduce **VastFSS**, covering 2,400 categories across diverse natural-image and specialized application contexts, providing a broader testbed for studying few-shot segmentation across a substantially larger category space.

Few-Shot Segmentation Methods. Recent FSS methods mainly follow prototype-based (Fan et al., 2022b; Liu et al., 2022; Okazawa, 2022; Wang et al., 2024; Zhang et al., 2021; Tian et al., 2020) or feature-matching-based (Hu et al., 2019; Zhang et al., 2022; Wang et al., 2020; Xie et al., 2021) paradigms, using support prototypes or dense support-query correspondences to guide segmentation. However, both methods rely largely on the limited visual evidence provided by support examples and may struggle with complex cases such as occlusions and visually similar distractors. More recently, MLLMs have been introduced to provide spatial (Zhu et al., 2024a) or semantic (Karimi & Poullis, 2025) guidance for FSS. These methods, however, often require auxiliary category names or external textual descriptions beyond the standard support image-mask input. In contrast, MA-FSS extracts target-aware representations directly from annotated support examples without requiring category names or external textual descriptions. It thus provides high-level semantic guidance while preserving the standard FSS input protocol.

3 The Proposed VastFSS

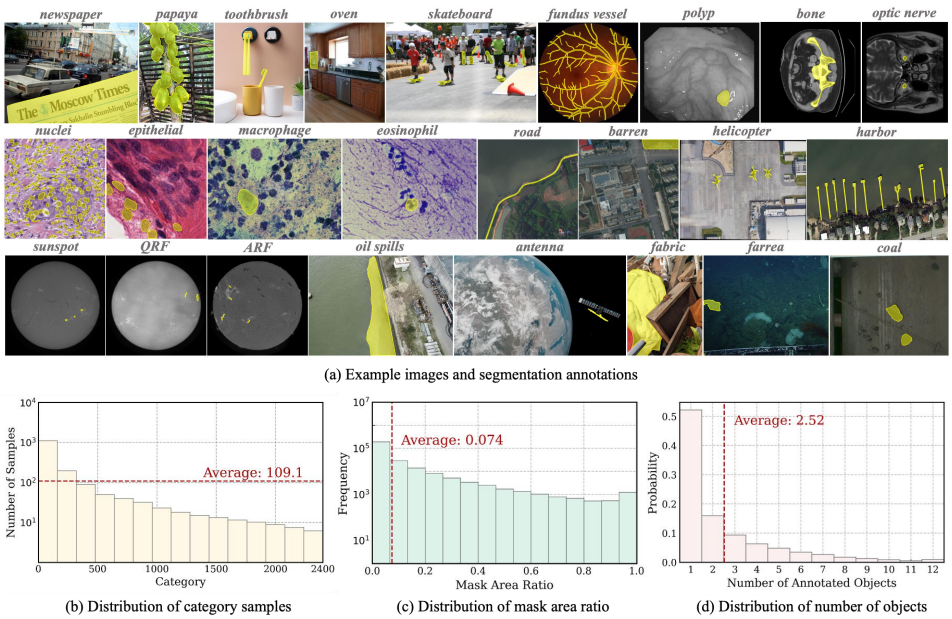
3.1 Construction Principle

VastFSS is designed to support the development and evaluation of generalizable FSS models through a large category vocabulary, broad coverage of visual concepts, and high-quality segmentation annotations. Its construction follows three principles. **(i) Large Category Vocabulary.** A benchmark for universal FSS should encompass a vast range of target categories. VastFSS therefore expands the category vocabulary available for training and evaluation, providing richer opportunities to learn transferable representations and assess adaptation to unseen categories from only a few annotated examples. **(ii) Broad Coverage of Visual Concepts.** Category diversity should extend beyond everyday objects in natural images. VastFSS incorporates targets from specialized applications, including medical imaging, remote sensing, biological imaging, and solar physics. This coverage complements the large vocabulary with varied target characteristics, imaging conditions, and segmentation challenges. **(iii) High-Quality Annotations.** High-quality mask annotations are essential for both training and evaluating few-shot segmentation models. To ensure annotation quality, VastFSS provides precise segmentation masks through careful inspection and refinement.

3.2 Data Acquisition

Category Construction. VastFSS is designed to cover a vast vocabulary of target concepts for few-shot segmentation. We select 2,400 categories from V3Det (Wang et al., 2023) and 59 datasets covering specialized applications, such as FIVES (Jin et al., 2022), FathomNet (Katija et al., 2022), and LoveDA (Wang et al., 2021). Each category is reviewed by a researcher with relevant expertise, such as a PhD or MS student working in related field, to assess its suitability for FSS. The selected categories extend beyond common natural-image concepts to include specialized targets with diverse visual characteristics and semantic definitions. The complete category inventory and corresponding data sources are provided in the ***appendix***.

Image Selection. After establishing the category inventory, we collect images from the corresponding source datasets. For categories derived from V3Det, we use the associated images from its detection dataset. For the remaining categories, we randomly sample an average of 60 images per category from the corresponding specialized datasets, covering varied imaging conditions and segmentation scenarios. This process yields an initial pool of 120K images. We then remove samples with ambiguous category correspondence, extremely small targets, severe occlusion, or poor visual quality. The resulting VastFSS benchmark comprises 104K images spanning 2,400 categories. Fig. 1 summarizes its statistics and compares it with existing FSS benchmarks. Further details on dataset maintenance, ethical considerations, and responsible use are provided in the ***appendix***.



(a) Example images and segmentation annotations

(b) Distribution of category samples

(c) Distribution of mask area ratio

(d) Distribution of number of objects

Figure 2: Overview and annotation statistics of VastFSS. (a) Examples of natural-image and specialized-domain categories, with ground-truth masks overlaid in yellow. (b) Distribution of samples per category. (c) Mask area ratio distribution. (d) Distribution of annotated objects per image.

3.3 Data Annotation

VastFSS provides pixel-level segmentation masks for all selected categories. For images from V3Det (Wang et al., 2023), we manually delineate target regions using the original bounding-box annotations as references for category identity and localization. For images selected from specialized datasets, we primarily use the corresponding segmentation annotations provided by the source datasets, as annotating these targets requires domain-specific expertise. These annotations are converted into the unified VastFSS format. Where necessary, missing masks are manually annotated and unsuitable masks are refined with reference to category descriptions and representative examples.

All masks undergo multiple rounds of verification by a validation team comprising two to three trained reviewers. Each annotation is assessed for semantic correctness and boundary accuracy to ensure that it covers the visible target region while excluding irrelevant regions. Annotations that fail to meet these criteria are corrected or re-annotated and reviewed again until they satisfy our quality requirements. The annotation pipeline is provided in the ***appendix***, and representative annotated examples from diverse application contexts are shown in Fig. 2 (a).

Statistics of annotation. To better understand VastFSS, we provide annotation statistics in Fig. 2, including the distributions of samples per category, mask area ratio, and annotated objects per image. As shown in Fig. 2 (b), the number of samples varies across categories and forms a long-tailed distribution, which reflects the imbalance commonly observed in real-world visual data and encourages the development of more practical and generalizable FSS models. Fig. 2 (c) shows that the average mask area ratio is 0.074, indicating that many targets occupy only a small portion of the image and thus require models to accurately segment specific targets from broad background regions. Fig. 2 (d) further reports the distribution of annotated objects per image, showing that VastFSS contains images with varying numbers of targets and provides diverse segmentation scenarios for evaluation.

3.4 Dataset Split and Evaluation Metric

Dataset Split. VastFSS contains 2,400 categories in total. Following Shaban et al. (2017); Nguyen & Todorovic (2019), we divide all categories into four disjoint folds, each containing 600 categories. During evaluation, one fold is used as the test set and the remaining three folds are used for training. This process is repeated four times so that each fold is used once for testing.

Tab. 1 summarizes the test-set statistics for the four splits. We aim to maintain broadly comparable category composition and numbers of images and annotated samples across splits. Each test fold covers a range of target concepts and application contexts, supporting evaluation of generalization to unseen categories. Since an image may contain categories from multiple folds, we assign images to non-overlapping train and test sets separately for each evaluation split. Images may recur across splits, so the image counts in Tab. 1 are not additive; 104K denotes the total number of unique images in VastFSS.

Table 1: Test-set statistics for the four evaluation splits of VastFSS. Image counts are reported separately for each split and are not additive across splits.

Test Fold	Test Categories	Test Images	Test Samples
Fold-0	600	29,997	40,653
Fold-1	600	30,098	40,624
Fold-2	600	29,294	40,270
Fold-3	600	30,151	40,957

For evaluation, VastFSS follows an episode-based protocol. In each episode, a target category is first sampled from the test fold, and then one query image and K (*e.g.*, one or five) annotated support images are sampled from this category. The model is required to segment the target category in the query image using only the support images and their masks as references.

Evaluation Metric. Following common practice in FSS (Shaban et al., 2017; Nguyen & Todorovic, 2019; Li et al., 2020), we use mean Intersection over Union (mIoU) as primary evaluation metric. For each test fold, IoU is computed between the predicted mask and the ground-truth for each query image. The category-level IoU is obtained by averaging over all query images of the same category, and the fold-level mIoU is then calculated by averaging IoU over all categories in the test fold.

4 MA-FSS: A New Method For VastFSS

4.1 Our MA-FSS Framework

Given a support image I_s and its corresponding mask M_s, we firstly construct a visual prompt by applying a red overlay to the annotated target region. The highlighted image is denoted by $\widetilde{I}_s = \texttt{Overlay}(I_s, M_s)$. The overlay explicitly identifies the target region while retaining its underlying visual content and surrounding context. We pair this image with a fixed textual instruction x_{txt}: "*This image is a support example. The target region is highlighted in red. Infer the visual concept represented by the highlighted region.*" The MLLM's image encoder extracts visual tokens from the highlighted support image, while the instruction is tokenized and mapped to text embeddings via $\mathbf{V} = \texttt{ImgEnc}(\widetilde{I}_s)$ and $\mathbf{T} = \texttt{Emb}(x_{\text{txt}})$, where $\mathbf{V}$ and $\mathbf{T}$ denote the visual and text tokens, respectively. Here, $\texttt{ImgEnc}(\cdot)$ denotes the MLLM's image encoding operation, and $\texttt{Emb}(\cdot)$ maps the tokenized instruction to text embeddings. The fixed instruction contains neither target category name nor any external target description. Thus, the target is specified solely by the support example, preserving the standard FSS input setting.

To extract compact target-aware representations, we introduce a sequence of N learnable tokens, $\mathcal{O} = [\mathbf{o}_1; \ldots; \mathbf{o}_N]$, to capture complementary characteristics of the target. We append them to the multimodal input after a special marker token, <token_start>, yielding $\mathbf{X} = [\mathbf{V}; \mathbf{T}; \mathbf{e}_{\text{start}}; \mathbf{o}_1; \ldots; \mathbf{o}_N]$,, where $\mathbf{e}_{\text{start}}$ is the embedding of <token_start>.

The MLLM processes the complete sequence in a single forward pass. The support mask M_s provides spatial guidance through the proposed mask-guided attention mechanism (Sec. 4.2), directing the learnable tokens toward the annotated foreground region. The resulting hidden states are

$$\mathbf{H} = \texttt{MLLM}(\mathbf{X}, M_s) \in \mathbb{R}^{L \times d}, \tag{1}$$

where L is the sequence length and d is the hidden dimension. We extract the hidden states at the learnable-token positions as follows,

$$\mathbf{h}_n = \mathbf{H}\,[\texttt{pos}(\mathbf{o}_n)] \in \mathbb{R}^d, \qquad n = 1, \ldots, N, \tag{2}$$

where $\texttt{pos}(\mathbf{o}_n)$ denotes the position of the n-th learnable token in the input sequence.

The resulting representations encode target information from the support image under the guidance of the instruction and the support mask. To adaptively adjust the contributions of individual tokens and their feature dimensions, we introduce an adaptive token gate (ATG, detailed in Sec. 4.3). ATG

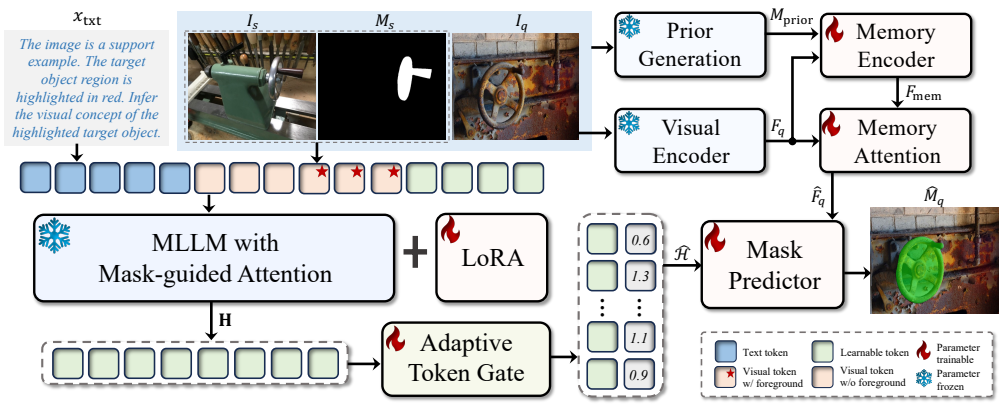


Figure 3: Overview of MA-FSS, which assists query segmentation with MLLM.

computes a gating vector for each token and applies it to produce the reweighted representations:

$$\widehat{\mathcal{H}} = [\widehat{\mathbf{h}}_1; \ldots; \widehat{\mathbf{h}}_N] = \texttt{ATG}([\mathbf{h}_1; \ldots; \mathbf{h}_N]) \in \mathbb{R}^{N \times d}, \tag{3}$$

where $\texttt{ATG}(\cdot)$ denotes the gating operation. This mechanism emphasizes relevant target information when constructing segmentation prompts.

While the learned tokens encode target characteristics, they do not explicitly provide spatial localization cues in the query image. We therefore complement these representations with a support-guided spatial prior. Following (Xu et al., 2025), we use a vision foundation model $\Phi_v(\cdot)$ (Oquab et al., 2024) to extract support and query features, denoted by $\widetilde{F}_s = \Phi_v(I_s)$ and $\widetilde{F}_q = \Phi_v(I_q)$, respectively. Then the support mask is used to construct foreground and background prototypes by pooling the masked features. Denoting these prototypes by F_{fg} and F_{bg}, we compute the spatial prior as

$$M_{\text{prior}} = \max\left(\texttt{CosSim}(F_{\text{fg}}, \widetilde{F}_q) - \texttt{CosSim}(F_{\text{bg}}, \widetilde{F}_q), 0\right), \tag{4}$$

where $\texttt{CosSim}(\cdot)$ computes cosine similarity between a prototype and the query feature at each spatial location. The resulting prior highlights query regions that are more similar to the support foreground than to its background. Further details of prior generation are provided in the ***appendix***.

To incorporate this spatial guidance, the visual encoder of SAM 3 (Carion et al., 2026) first extracts query features $F_q = \mathcal{E}_V(I_q)$. Its memory encoder then combines these features with the spatial prior mask M_{prior} to construct a target memory representation. Memory attention subsequently integrates this representation into the query features, yielding

$$F_{\text{mem}} = \texttt{MemEnc}(F_q, M_{\text{prior}}), \quad \widehat{F}_q = \texttt{MemAtten}(F_q, F_{\text{mem}}), \tag{5}$$

where $\texttt{MemEnc}(\cdot)$ and $\texttt{MemAtten}(\cdot)$ denote the memory encoder and memory attention module. The resulting query features incorporate the spatial guidance provided by the support sample.

Finally, the mask predictor combines the reweighted target tokens $\widehat{\mathcal{H}}$ with the memory-enhanced query features $\widehat{F}_q$ to predict the query mask,

$$\widehat{M}_q = \texttt{MaskPred}(\widehat{F}_q, \widehat{\mathcal{H}}). \tag{6}$$

Here $\texttt{MaskPred}(\cdot)$ refers to the mask predictor of SAM 3. Together, the target-aware tokens and the spatial prior provide complementary guidance for query segmentation.

4.2 Mask-Guided Attention Mechanism

To help the learnable tokens extract target-aware representations while reducing interference from background distractors, we introduce a mask-guided attention mechanism. It uses the support mask to restrict direct attention from learnable tokens to visual tokens associated with the annotated foreground region. The learnable tokens can also attend to the textual instruction and other learnable tokens, allowing them to incorporate task context and exchange complementary target information.

Let $\mathcal{I}_{\text{vision}}$, $\mathcal{I}_{\text{text}}$ and $\mathcal{I}_{\text{learn}}$ denote the index sets of visual, text and learnable tokens, respectively. We first spatially align the binary support mask M_s with the visual-token grid and partition it into the corresponding patches. Let $\widetilde{M}_s(j)$ denote the mask patch associated with the j-th visual token. The foreground visual-token indices are defined as

$$\mathcal{I}_{\text{fg}} = \left\{ j \in \mathcal{I}_{\text{vision}} \;\middle|\; \texttt{sum}(\widetilde{M}_s(j)) > 0 \right\}, \tag{7}$$

where $\texttt{sum}(\widetilde{M}_s(j))$ is the sum of the binary mask values within the corresponding patch. Thus, a visual token is considered a foreground token if its spatial region overlaps the annotated target. Let $\mathbf{A} \in \{0,1\}^{L\times L}$ denote the attention mask, where $\mathbf{A}_{ij} = 1$ indicates that the token at position i may attend to the token at position j. We preserve the original causal attention pattern for visual and text tokens. For learnable tokens, we allow attention to foreground visual tokens, all text tokens, and all learnable tokens, while blocking direct attention to background visual tokens. As illustrated in Fig. 4, the attention mask is defined as

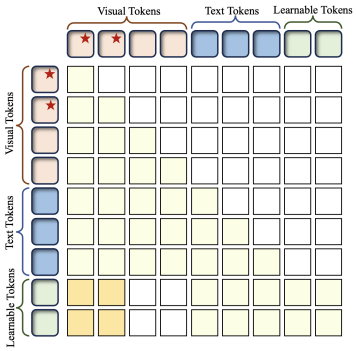


Figure 4: Illustration of the mask-guided attention.

$$\mathbf{A}_{ij} = \begin{cases} 1, & i \in \mathcal{I}_{\text{vision}} \cup \mathcal{I}_{\text{text}}, \quad j \le i, \\ 1, & i \in \mathcal{I}_{\text{learn}}, \quad j \in \mathcal{I}_{\text{fg}} \cup \mathcal{I}_{\text{text}} \cup \mathcal{I}_{\text{learn}}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

This attention pattern directs the learnable tokens toward foreground visual information while preserving access to textual instruction and enabling bidirectional interaction among the learnable tokens. By excluding background visual tokens from attention, the mechanism reduces background interference and encourages target-aware representations for the mask prediction.

4.3 Adaptive Token Gate

Although the target-aware tokens capture complementary characteristics of the target, they may also encode redundant or less relevant information. We therefore introduce an adaptive token gate (ATG) to modulate their contributions at the feature level.

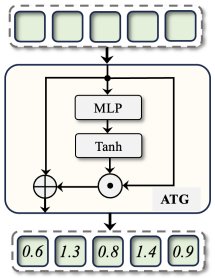


Figure 5: Overview of ATG.

As illustrated in Fig. 5, given the target-aware representations $\{\mathbf{h}_n\}_{n=1}^{N}$ produced by the MLLM, a shared multilayer perceptron (MLP) predicts a modulation vector for each token. An element-wise hyperbolic tangent activation bounds its values as follows,

$$\mathbf{w}_n = \texttt{Tanh}\left(\texttt{MLP}(\mathbf{h}_n)\right) \in (-1,1)^d, \qquad n = 1,\ldots,N, \tag{9}$$

where $\mathbf{w}_n$ contains the modulation coefficients for feature dimensions of the n-th token. We then apply residual gating to obtain the reweighted representation:

$$\widehat{\mathbf{h}}_n = \mathbf{h}_n + 0.5\,\mathbf{w}_n \odot \mathbf{h}_n, \qquad n = 1,\ldots,N, \tag{10}$$

where $\odot$ denotes element-wise multiplication. Each effective scaling coefficient lies in $(0.5, 1.5)$, allowing ATG to attenuate less relevant features and amplify informative ones while retaining a residual path for the original token representation. The resulting reweighted tokens are used as target-aware semantic prompts for query mask prediction.

4.4 K-shot Extension and Optimization

Extending to K-shot Setting. In the K-shot setting, each annotated support sample produces N target-aware token representations. We aggregate them by averaging corresponding token positions across the K support samples, $\overline{\mathbf{h}}_j = \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathbf{h}}_j^{(k)}$, for $j = 1,\ldots,N$, where $\widehat{\mathbf{h}}_j^{(k)}$ denotes the j-th reweighted concept representation extracted from the k-th support sample. The resulting N representations serve as prompts for the segmentation model. Similarly, each support sample is

Table 2: Comparisons with state-of-the-arts on VastFSS. Best results are highlighted in **bold**.

Method	1-shot					5-shot				
	Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
DiffewS (Zhu et al., 2024b) [NeurIPS'24]	43.5	46.8	47.0	46.0	45.8	51.5	52.4	53.3	52.6	52.5
GF-SAM (Zhang et al., 2024a) [NeurIPS'24]	35.8	35.9	36.1	37.9	36.4	44.4	43.7	44.4	43.9	44.1
Matcher (Liu et al., 2024b) [ICLR'24]	27.0	28.2	28.8	28.1	28.0	27.3	30.3	30.2	28.5	29.1
VRP-SAM (Sun et al., 2024a) [CVPR'24]	36.1	38.4	38.3	35.1	37.0	43.6	45.8	46.3	43.1	44.7
FCP (Park et al., 2025) [AAAI'25]	37.4	39.0	38.0	36.8	37.8	44.5	46.7	45.9	46.0	45.8
FSSAM (Xu et al., 2025) [ICML'25]	42.1	44.9	42.2	43.8	43.3	52.9	55.4	54.2	54.4	54.2
UNICL-SAM (Sheng et al., 2025) [CVPR'25]	42.6	45.4	38.1	40.3	41.6	45.8	51.6	52.1	45.5	48.8
SANSA (Cuttano et al., 2025) [NeurIPS'25]	52.7	52.5	52.6	49.5	51.8	54.8	56.8	57.1	56.2	56.2
PR-MaGIC (Lee et al., 2026) [CVPR'26]	31.4	32.2	31.8	31.5	31.7	33.6	34.4	34.4	35.6	34.5
INSID3 (Cuttano et al., 2026) [CVPR'26]	38.0	38.1	37.6	37.1	37.7	44.5	45.3	45.1	46.3	45.3
MA-FSS (ours)	**54.0**	**56.9**	**54.4**	**55.8**	**55.3**	**56.6**	**59.3**	**58.7**	**57.5**	**58.0**

paired with the query image to generate a spatial prior $M_{\text{prior}}^{(k)}$. These priors are averaged to obtain $M_{\text{prior}} = \frac{1}{K}\sum_{k=1}^{K} M_{\text{prior}}^{(k)}$. The subsequent memory encoding, memory attention, and mask prediction follow the same procedure as in the 1-shot setting.

Optimization. Given a query image and an annotated support set, MA-FSS is jointly optimized using an autoregressive text generation loss and a segmentation loss. Due to space limitation, please refer to the ***appendix*** for details of the loss function.

5 Experiments

Implementation. MA-FSS is implemented in PyTorch (Paszke et al., 2019) with Qwen3-VL-2B (Bai et al., 2025) as the MLLM and SAM 3 (Carion et al., 2026) for segmentation. We use $N = 10$ learnable tokens and DINOv2-B (Oquab et al., 2024) for prior generation. LoRA with rank 256 is applied to the MLLM, while the SAM 3 memory encoder, memory attention module, and mask predictor are fully fine-tuned; newly introduced modules are trained from scratch. Training is conducted for 80,000 iterations on two NVIDIA A800 GPUs using DeepSpeed (Rasley et al., 2020), with an effective batch size of 4. We use AdamW (Loshchilov & Hutter, 2019) with learning rates of 5×10^{-6} for the MLLM and 1×10^{-5} for the remaining trainable modules. The 1-shot model is directly used for 5-shot evaluation without additional fine-tuning. Unless otherwise specified, results are averaged over 2,000 randomly sampled episodes.

5.1 State-of-the-art Comparison

We compare MA-FSS with recent state-of-the-art FSS methods on the proposed VastFSS benchmark. As reported in Tab. 2, MA-FSS consistently outperforms the competing methods across all four folds under both the 1-shot and 5-shot settings, achieving mean mIoU scores of 55.3% and 58.0%. These results demonstrate its effectiveness in learning target-aware representations from limited support examples and generalizing to unseen categories across diverse visual domains.

5.2 Ablation Study

Ablation Study on MA and ATG. We evaluate the contributions of mask-guided attention (MA) and the adaptive token gate (ATG) in Tab. 3. The baseline uses standard causal attention without either component and achieves a mean mIoU of 53.6% (❶). Introducing MA increases the score to 54.6% (❷), demonstrating the benefit of directing learnable tokens toward the annotated foreground region. Adding ATG further improves performance to 55.3% (❸), yielding an overall gain of 1.7 percentage points over the baseline. These results support the combined effectiveness of foreground-guided attention and adaptive token reweighting.

Comparison with Alternative Token Combination Strategies. We compare ATG with two alternative combination strategies in Tab. 4. Average pooling combines all learnable tokens into a single representation with equal weights (❶), whereas concatenation followed by an MLP produces a learned fused representation (❷). In contrast, ATG retains individual tokens and adaptively modulates their features (❸). ATG achieves the best performance across all four folds, with a mean mIoU of 55.3%, compared with 54.1% for average pooling and 52.4% for concatenation. These re-

Table 3: Ablation on the mask-guided attention mechanism (MA) and ATG.

	MA	ATG	Fold-0	Fold-1	Fold-2	Fold-3	Mean
❶	-	-	52.3	53.7	52.9	55.5	53.6
❷	✓	-	53.7	55.0	54.1	55.7	54.6
❸	✓	✓	**54.0**	**56.9**	**54.4**	**55.8**	**55.3**

Table 4: Ablation on the comparison of ATG with other combination strategies.

	Strategy	Fold-0	Fold-1	Fold-2	Fold-3	Mean
❶	Addition	52.0	55.1	54.1	55.1	54.1
❷	Concatenation	51.8	50.4	52.9	54.3	52.4
❸	ATG (ours)	**54.0**	**56.9**	**54.4**	**55.8**	**55.3**

sults suggest that retaining multiple target-aware representations with adaptive reweighting is more effective than compressing them into a single representation.

Ablation on the Number of Learnable Tokens. We investigate the effect of the number of learnable tokens by varying $N \in \{1, 5, 10, 15\}$. As shown in Tab. 5, performance improves as N increases from 1 to 10, suggesting that additional learnable tokens help capture complementary target information. However, increasing N to 15 leads to a performance decline (❹), indicating that more tokens do not necessarily yield better representations and may introduce redundancy. We therefore set $N = 10$ by default (❸), as it achieves the best performance among the evaluated settings.

Table 5: Ablation on the number N.

	N	Fold-0	Fold-1	Fold-2	Fold-3	Mean
❶	1	50.1	56.0	54.3	55.5	54.0
❷	5	53.8	56.8	54.1	54.7	54.9
❸	10	**54.0**	**56.9**	**54.4**	**55.8**	**55.3**
❹	15	**54.0**	55.8	54.2	54.8	54.7

5.3 Zero-shot Transfer

Table 6: Comparisons on COCO-20^i. Here "20^i" denotes the mIoU score of the i-th fold.

Method	20^0	20^1	20^2	20^3	Mean
VRP-SAM (Sun et al., 2024a) [CVPR'24]	56.8	61.0	64.2	59.7	60.4
DeFSS (Qin et al., 2025) [ICCV'25]	46.5	59.7	54.8	53.2	53.7
FSSAM (Xu et al., 2025) [ICML'25]	59.9	65.6	62.1	61.6	62.3
SANSA (Cuttano et al., 2025) [NeurIPS'25]	58.9	62.6	61.5	58.0	60.2
BPNet (Shi et al., 2026) [CVPR'26]	46.7	59.2	56.8	54.2	54.2
INSID3 (Cuttano et al., 2026) [CVPR'26]	53.9	60.4	56.3	55.1	56.4
MA-FSS (zero-shot transfer)	57.6	**71.6**	62.9	64.6	64.2
MA-FSS (trained on COCO-20^i)	**62.3**	67.3	**66.4**	**69.0**	**66.3**

Table 7: Comparison on Chest X-ray and FSS-1000 under the 1-shot setting.

Method	Chest X-ray	FSS-1000
APM (Tong et al., 2024) [NeurIPS'24]	78.3	79.3
IFA (Nie et al., 2024) [CVPR'24]	74.0	80.1
SDRC (Tong et al., 2025b) [ICML'25]	82.9	80.3
DFN (Tong et al., 2025a) [ICML'25]	83.2	83.0
LoEC (Liu et al., 2025a) [CVPR'25]	83.9	81.1
HERA (Ma et al., 2026) [CVPR'26]	85.8	81.6
DHANet (Sun et al., 2026) [ECCV'26]	86.1	83.6
MA-FSS (zero-shot transfer)	**87.7**	**89.4**

To assess its effectiveness beyond VastFSS, we further evaluate MA-FSS on COCO-20^i under the 1-shot setting. As shown in Tab. 6, MA-FSS achieves a mean mIoU of 66.3%, the highest among the compared methods. We also directly evaluate the VastFSS-trained model on COCO-20^i without additional training or fine-tuning. This model achieves a mean mIoU of 64.3%, remaining competitive with recent FSS methods. These results provide evidence of the transferability of MA-FSS and the utility of VastFSS as a training benchmark.

We further evaluate the VastFSS-trained model under the 1-shot setting on two datasets used in the CD-FSS benchmark, namely Chest X-ray (Candemir et al., 2013) and FSS-1000 (Li et al., 2020), covering lung segmentation in chest radiographs and segmentation across a broad range of object categories in natural images. As reported in Tab. 7, MA-FSS achieves the highest mIoU among the compared methods on both datasets, reaching 87.7% on Chest X-ray and 89.4% on FSS-1000. These results demonstrate its competitive performance against methods specifically designed for cross-domain FSS and further support its effectiveness across different segmentation applications.

6 Conclusion

In this work, we introduce VastFSS, a large-vocabulary benchmark designed to advance universal few-shot segmentation. VastFSS covers 2,400 target categories across a wide range of application scenarios, broadening the scope of FSS training and evaluation. We further propose MA-FSS, a simple yet effective baseline that combines MLLM-derived target representations with a support-guided spatial prior. Extensive experiments demonstrate that MA-FSS outperforms recent FSS methods on VastFSS and transfers effectively to existing FSS benchmarks without additional training or fine-tuning. These results validate the effectiveness of MA-FSS and highlight the potential of VastFSS to support future research toward universal few-shot segmentation.

AI USE STATEMENT

Generative AI tools were used solely for brainstorming and language polishing. They were not used to generate data or experimental results, formulate mathematical claims, design experiments, implement methods, analyze results, or interpret findings. All AI-assisted text was carefully reviewed, revised, and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

VastFSS is constructed from publicly available datasets. We document and acknowledge all data sources and adhere to their respective licenses and terms of use throughout data collection, processing, and integration. VastFSS and MA-FSS are intended for research and have not been validated for clinical decision-making or other safety-critical applications. Deployment in such settings requires further domain-specific validation and appropriate human oversight. We discourage any use that violates privacy or facilitates discriminatory surveillance. We provide more detailed ethical statement in the ***appendix***.

REFERENCES

Iftekharul Abedeen, Md. Ashiqur Rahman, Fatema Zohra Prottyasha, Tasnim Ahmed, Tareque Mohmud Chowdhury, and Swakkhar Shatabda. Fracatlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs. *Scientific Data*, 10(1), 2023.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28, 2020.

Oluwatosin Alabi, Ko Ko Zayar Toe, Zijian Zhou, Charlie Budd, Nicholas Raison, Miaojing Shi, and Tom Vercauteren. Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery. *Scientific Data*, 12(1), 2025.

Kooshan Amini, Yuhao Liu, Jamie Ellen Padgett, Guha Balakrishnan, and Ashok Veeraraghavan. Debris segmentation using post-hurricane aerial imagery. *Computer-Aided Civil and Infrastructure Engineering*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. In *arXiv*, 2025.

Hanbo Bi, Yingchao Feng, Yongqiang Mao, Jianning Pei, Wenhui Diao, Hongqi Wang, and Xian Sun. Agmtr: Agent mining transformer for few-shot segmentation in remote sensing. In *IJCV*, 2025.

Kyle Bradbury, Raghav Saboo, Timothy L. Johnson, Jordan M. Malof, Arjun Devarajan, Wuming Zhang, Leslie M. Collins, and Richard G. Newell. Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. *Scientific Data*, 3, 2016.

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2), 2013.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *ICLR*, 2026.

Maren Carstens, Fiona R Kolbinger, Sebastian Bodenstedt, Alexander C Jenke, Marius Distler, Jürgen Weitz, and Stefanie Speidel. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1), 2023.

Priti Chakurkar and Deepali Vora. Context-aware deep learning based indian footpath damage segmentation dataset for risk assessment. *Scientific Data*, 12(1), 2025.

Gaowen Chen, Jieyun Bai, Zhanhong Ou, Yaosheng Lu, and Huijin Wang. Psfhs: intrapartum ultrasound image dataset for ai-based segmentation of pubic symphysis and fetal head. *Scientific Data*, 11(1), 2024.

Mang Tik Chiu, Xingqian Xu, Kai Wang, Jennifer Hobbs, Naira Hovakimyan, Thomas S Huang, and Honghui Shi. The 1st agriculture-vision challenge: Methods and results. In *CVPRW*, 2020.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). In *arXiv*, 2019.

Claudia Cuttano, Gabriele Trivigno, Giuseppe Averta, and Carlo Masone. Sansa: Unleashing the hidden semantics in sam2 for few-shot segmentation. In *NeurIPS*, 2025.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. Insid3: Training-free in-context segmentation with dinov3. In *CVPR*, 2026.

Thomas De Kerf, Seppe Sels, Svetlana Samsonova, and Steve Vanlanduit. A dataset of drone-captured, segmented images for oil spill detection in port environments. *Scientific Data*, 11, 2024.

Antun Duras, Ben J. Wolf, Athina Ilioudi, Ivana Palunko, Bart De Schutter, and Robert Babuska. A dataset for detection and segmentation of underwater marine debris in shallow waters. *Scientific Data*, 11(1), 2024.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 2022a.

Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *ECCV*, 2022b.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification. In *ECDP*, 2019.

Omid Ghorbanzadeh, Yonghao Xu, Hengwei Zhao, Junjue Wang, Yanfei Zhong, Dong Zhao, Qi Zang, Shuang Wang, Fahong Zhang, Yilei Shi, Xiao Xiang Zhu, Lin Bai, Weile Li, Weihang Peng, and Pedram Ghamisi. The outcome of the 2022 landslide4sense competition: Advanced landslide detection from multisource satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 2022.

Ke Gong, Xiaodan Liang, Dongyu Li, Yichen Wang, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017.

Francesco Guarnera, Alessia Rondinella, Elena Crispino, Giulia Russo, Clara Di Lorenzo, Davide Maimone, Francesco Pappalardo, and Sebastiano Battiato. Mslesseg: baseline and benchmarking of a new multiple sclerosis lesion segmentation dataset. *Scientific Data*, 12(1), 2025.

Ruoyi Hao, Yang Zhang, Zhiqing Tang, Yang Zhou, Lalithkumar Seenivasan, Catherine Po Ling Chan, Jason Ying Kuen Chan, and Hongliang Ren. Upper airway anatomical landmark dataset for automated bronchoscopy and intubation. *Scientific Data*, 12(1), 2025.

Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees GM Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *AAAI*, 2019.

Lixuan Huang, Jiangnian Gong, Daqin Feng, Ling Zhang, Hao Ren, Xin Zhao, Chang Liu, Hui Liang, Panlin Mo, Minhai Dong, Yongjia Yu, Zisan Zeng, and Lun Liang. A comprehensive dataset of germinoma on mri/ct with clinical and radiomic data. *Scientific Data*, 12(1), 2025.

Shiqi Huang, Tingfa Xu, Ning Shen, Feng Mu, and Jianan Li. Rethinking few-shot medical segmentation: a vector quantization view. In *CVPR*, 2023.

Debesh Jha, Michael H Smedsrud, Michael A Riegler, Debora L Johansen, Pål Halvorsen, Håvard D Johansen, Thomas de Lange, and Peter T Schmidt. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling*, 2020.

Kai Jin, Xingru Huang, Jing Zhou, Yunxiang Li, Yifan Yan, Yalin Miao, Qiang Zhang, and Juan Ye. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 9(1), 2022.

Hossein Kabir, Jordan Wu, Sunav Dahal, Tony Joo, and Nishant Garg. Sorpvision: A comprehensive dataset for cementitious sorptivity analysis powered by computer vision. *Scientific Data*, 12(1), 2025.

Jian Kang, Fengyu Tong, Xiang Ding, Sijiang Li, Ruoxin Zhu, Yan Huang, Yusheng Xu, and Rubén Fernández-Beltran. Decoding the partial pretrained networks for sea-ice segmentation of 2021 gaofen challenge. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022.

Amin Karimi and Charalambos Poullis. Dsv-lfs: Unifying llm-driven semantic cues with visual features for robust few-shot segmentation. In *CVPR*, 2025.

Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, and Katherine L. C. Bell. Fathomnet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports*, 12, 2022.

A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical image analysis*, 69:101950, 2021.

Rana Khaled, Maha Helal, Omar Alfarghaly, Omnia Mokhtar, Abeer Elkorany, Hebatalla El Kassas, and Aly Fahmy. Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. *Scientific Data*, 9, 2022.

Jessie Kulaga-Yoskovitz, Boris C. Bernhardt, Seok-Jun Hong, Tommaso Mansi, KE Liang, Andrea Bernasconi, and Neda Bernasconi. Multi-contrast submillimetric 3 tesla hippocampal subfield segmentation protocol and dataset. *Scientific Data*, 2(1), 2015.

Neeraj Kumar, Ruchika Verma, Sanket Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7), 2017.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.

Bennett Landman, Zhipeng Xu, J. Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault Workshop Challenge*, volume 5, 2015.

Minjae Lee, Sungwoo Hur, Soojin Hwang, and Won Hwa Kim. Pr-magic: Prompt refinement via mask decoder gradient flow for in-context segmentation. In *CVPR*, 2026.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *ECCV*, 2022.

Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, 2020.

Yin Li, Qi Chen, Kai Wang, Meige Li, Liping Si, Yingwei Guo, Yu Xiong, Qixing Wang, Yang Qin, Ling Xu, Patrick van der Smagt, Jun Tang, and Nutan Chen. A dataset of primary nasopharyngeal carcinoma mri with multi-modalities segmentation. *Scientific Data*, 12(1), 2025.

S. Lin, L. Huang, X. Liu, G. Chen, and Z. Fu. A construction waste landfill dataset of two districts in beijing, china from high resolution satellite images. *Scientific Data*, 11(1), 2024.

Peng Liu, Svea Steuer, Jonas Golde, Joseph Morgenstern, Yujia Hu, Clara Schieffer, Sven Oßmann, Marcus Neudert, Edmund Koch, Stefanie Speidel, et al. The dresden in vivo oct dataset for automatic middle ear segmentation. *Scientific Data*, 11(1), 2024a.

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 *ICLR*, 2024b.

Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *CVPR*, 2022.

Yuhan Liu, Yixiong Zou, Yuhua Li, and Ruixuan Li. The devil is in low-level features for cross-domain few-shot segmentation. In *CVPR*, 2025a.

Yuxuan Liu, Chunjiang Bian, Hongbin Nie, Shi Chen, and Ziqian Yang. A large-scale synthetic benchmark dataset for non-cooperative space target perception. *Scientific Data*, 12(1), 2025b.

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

Ziqi Lv, Yuhan Fan, Jie Sha, et al. A large-scale open image dataset for deep learning-enabled intelligent sorting and analyzing of raw coal. *Scientific Data*, 12(1), 2025.

Junyuan Ma, Xunzhi Xiang, Wenbin Li, Qi Fan, and Yang Gao. Selective, regularized, and calibrated: Harnessing vision foundation models for cross-domain few-shot semantic segmentation. *CVPR*, 2026.

Pietro Mascagni, Adithya Murali, Deepak Alapatt, Armine Vardazaryan, A. Garcia, C. I. Nwoye, E. Flouty, C. Gonzalez, T. Yu, A. Vardazaryan, and Nicolas Padoy. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data*, 12(1), 2025.

S. Mohajerani and P. Saeedi. Cloud-net: An end-to-end cloud detection algorithm for landsat 8 imagery. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019.

Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 2016.

Roberto Morelli, Luca Clissa, Roberto Amici, Matteo Cerri, Timna Hitrec, Marco Luppi, Lorenzo Rinaldi, Fabio Squarcio, and Antonio Zoccoli. Automating cell counting in fluorescent microscopy through deep learning with c-resunet. *Scientific Reports*, 11(1), 2021.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, 2019.

Jiahao Nie, Yun Xing, Gongjie Zhang, Pei Yan, Aoran Xiao, Yap-Peng Tan, Alex C Kot, and Shijian Lu. Cross-domain few-shot segmentation via iterative support-query correspondence mining. In *CVPR*, 2024.

Atsuro Okazawa. Interclass prototype relation for few-shot segmentation. In *ECCV*, 2022.

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. In *TMLR*, 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 *ECCV*, 2020.

Danielle F Pace et al. Hvsmr-2.0: A 3d cardiovascular mr dataset for whole-heart segmentation in congenital heart disease. *Scientific Data*, 11(1), 2024.

Suho Park, SuBeen Lee, Hyun Seok Seong, Jaejoon Yoo, and Jae-Pil Heo. Foreground-covering prototype generation and matching for sam-aided few-shot segmentation. In *AAAI*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, Hui Ji, Levente Lanczi, Marianna Nagy, Monika Beresova, Thi Dao Nguyen, Giancarlo Natalucci, Theofanis Karayannis, Bjoern Menze, Meritxell Bach Cuadra, and Andras Jakab. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific Data*, 8(1), 2021.

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, 2015.

Prasanna Porwal, Samiksha Pachade, Rohan Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3), 2018.

Guoqiang Qi, Xiongfei Jiao, Jing Li, Chaojin Qin, Xinxin Li, Zhexian Sun, Yonggen Zhao, Renjie Jiang, Zhu Zhu, Guoqiang Zhao, et al. A dataset for quality evaluation of pelvic x-ray and diagnosis of developmental dysplasia of the hip. *Scientific Data*, 12(1), 2025.

Zishu Qin, Junhao Xu, and Weifeng Ge. Defss: Image-to-mask denoising learning for few-shot segmentation. In *ICCV*, 2025.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, 2020.

Antonio Reche and Consuelo Cid. Solar filament detection, classification, and tracking with deep learning. In *Proceedings of SPAICE2024: The First Joint European Space Agency/IAA Conference on AI in and for Space*, 2024.

Bodhisattwa Rister, David Daraby, Markus Kowarschik, Rosemarie Do, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7 (1), 2020.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017.

Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Tao Gong, Bin Liu, Jing Han, Wenbin Tu, Shengwei Xu, et al. Unicl-sam: Uncertainty-driven in-context segmentation with part prototype discovery. In *CVPR*, 2025.

Guangchen Shi, Yirui Wu, Wei Zhu, Tao Wang, Hao Zhang, Bo Li, and Tong Lu. Bayesian decomposition and semantic completion for few-shot semantic segmentation. In *CVPR*, 2026.

Xiaomei Shi, Qing Huang, Tian Xu, Heng Mei, Tingwei Quan, and Xi Wang. Pw-balfc, a clinical dataset for detection and instance segmentation of bronchoalveolar lavage fluid cell. *Scientific Data*, 12(1), 2025.

D. Sirimewan, S. Dayarathna, S. Raman, and Y. Bai. A benchmark dataset for class-wise segmentation of construction and demolition waste in cluttered environments. *Scientific Data*, 12(1), 2025.

Sujun Sun, Mingwu Ren, and Haofeng Zhang. Hierarchical spatial and channel aggregation for cross-domain few-shot segmentation. In *ECCV*, 2026.

Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. Vrp-sam: Sam with visual reference prompt. In *CVPR*, 2024a.

Yiming Sun, Nuliqiman Maimaiti, Peifang Xu, Peng Jin, Jingxuan Cai, Guiping Qian, Pengjie Chen, Mingyu Xu, Gangyong Jia, Qing Wu, et al. An as-oct image dataset for deep learning-enabled segmentation and 3d reconstruction for keratitis. *Scientific Data*, 11, 2024b.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3), 2020.

Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2), 2020.

Jintao Tong, Yixiong Zou, Yuhua Li, and Ruixuan Li. Lightweight frequency masker for cross-domain few-shot semantic segmentation. In *NeurIPS*, 2024.

Jintao Tong, Ran Ma, Yixiong Zou, Guangyao Chen, Yuhua Li, and Ruixuan Li. Adapter naturally serves as decoupler for cross-domain few-shot semantic segmentation. In *ICML*, 2025a.

Jintao Tong, Yixiong Zou, Guangyao Chen, Yuhua Li, and Ruixuan Li. Self-disentanglement and re-composition for cross-domain few-shot segmentation. In *ICML*, 2025b.

Noelia Vallez, Gloria Bueno, Oscar Deniz, Miguel Angel Rienda, and Carlos Pastor. Bus-uclm: Breast ultrasound lesion segmentation dataset. *Scientific Data*, 12, 2025.

Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *ECCV*, 2020.

Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *ICCV*, 2023.

Jing Wang, Jiangyun Li, Chen Chen, Yisi Zhang, Haoran Shen, and Tianxiang Zhang. Adaptive fss: a novel few-shot segmentation framework via prototype enhancement. In *AAAI*, 2024.

Junjue Wang, Zhuo Zheng, AilongX Ma, Xiaoyang Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *NeurIPS*, 2021.

Chunlin Wen, Yu Zhang, Jie Fan, Hongyuan Zhu, Xiu-Shen Wei, Yijun Wang, Zhiqiang Kou, and Shuzhou Sun. Object-level correlation for few-shot segmentation. In *ICCV*, 2025.

Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *CVPR*, 2021.

Qianxiong Xu, Guosheng Lin, Chen Change Loy, Cheng Long, Ziyue Li, and Rui Zhao. Eliminating feature ambiguity for few-shot segmentation. In *ECCV*, 2024a.

Qianxiong Xu, Lanyun Zhu, Xuanyi Liu, Guosheng Lin, Cheng Long, Ziyue Li, and Rui Zhao. Unlocking the power of sam 2 for few-shot segmentation. In *ICML*, 2025.

Zishan Xu, Yuqing Yang, Qianzhen Fang, Wei Chen, Tingting Xu, Jueting Liu, and Zehua Wang. A comprehensive dataset for digital restoration of dunhuang murals. *Scientific Data*, 11, 2024b.

Shunhan Yao, Yuanxiang Huang, Xiaoyu Wang, Yiwen Zhang, Ian Costa Paixao, Zhikang Wang, Charla Lu Chai, Hongtao Wang, Dinggui Lu, Geoffrey I Webb, et al. A radiograph dataset for the classification, localization, and segmentation of primary bone tumors. *Scientific Data*, 12, 2025.

Mohammed Yaqoob, Mohammed Ishaq, Mohammed Yusuf Ansari, Venkata Ram Sagar Konagandla, Tamim Al Tamimi, Stefano Tavani, Amerigo Corradetti, and Thomas Daniel Seers. Geocrack: A high-resolution dataset for segmentation of fracture edges in geological outcrops. *Scientific Data*, 11, 2024.

Xin Ye, Shucheng He, Xiaxing Zhong, Jiafeng Yu, Shangchao Yang, Yingjiao Shen, Yiqi Chen, Yaqi Wang, Xingru Huang, and Lijun Shen. Oimhs: An optical coherence tomography image dataset based on macular hole manual segmentation. *Scientific Data*, 10(1), 2023.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with mllm for dense grounded understanding of images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Ling Shao Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *CVPRW*, 2019.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *NeurIPS*, 2024a.

Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *CVPR*, 2021.

Gengwei Zhang, Shant Navasardyan, Ling Chen, Yao Zhao, Yunchao Wei, Humphrey Shi, et al. Mask matching transformer for few-shot segmentation. In *NeurIPS*, 2022.

Haiyang Zhang, Hoi Chi Chan, Jiashuo Xu, Mengda Jiang, Xiaofeng Tao, Huifang Zhou, Xuefei Song, and Xianqun Fan. Tom500: A multi-organ annotated orbital mri dataset for thyroid eye disease. *Scientific Data*, 12, 2025.

Yichi Zhang, Jing Wang, Tan Pan, Quanling Jiang, Jingjie Ge, Xin Guo, Chen Jiang, Jie Lu, Jianning Zhang, Xueling Liu, Mei Tian, Yuan Qi, Yuan Cheng, and Chuantao Zuo. Nasalseg: A dataset for automatic segmentation of nasal cavity and paranasal sinuses from 3d ct images. *Scientific Data*, 11(1), 2024b.

Yifan Zhang, Fan Ye, Lingxiao Chen, Feng Xu, Xiaodiao Chen, Hongkun Wu, Mingguo Cao, Yunxiang Li, Yaqi Wang, and Xingru Huang. Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. *Scientific Data*, 10:366, 2023.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *ECCV*, 2024c.

Z. Zhong et al. A comprehensive dataset of magnetic resonance enterography images with intestinal segment annotations. *Scientific Data*, 12(1), 2025.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large language models meet few-shot segmentation. In *CVPR*, 2024a.

Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. In *NeurIPS*, 2024b.

Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018.

Tianyu Zou, Shengwu Xiong, Ruilin Yao, and Yi Rong. Balancing conservatism and aggressiveness: Prototype-affinity hybrid network for few-shot segmentation. In *ICCV*, 2025.

APPENDIX

This appendix provides additional details on the construction of VastFSS, the MA-FSS framework, and our experimental evaluation. It is organized as follows.

- **A Category Inventory and Data Sources**
  We provide the complete category inventory of VastFSS and document the corresponding data sources.
- **B Construction Pipeline and Additional Annotation Examples**
  We describe the dataset construction pipeline and present additional annotation examples.
- **C Details of the Prior Generation Module**
  We detail the construction of support-guided spatial prior mask.
- **D Details of Optimization**
  We describe the training objectives and mask selection procedure.
- **E Qualitative Analysis**
  We visualize and compare segmentation results on VastFSS to illustrate the effectiveness of MA-FSS.
- **F Additional Results on LVIS-92^i**
  We compare MA-FSS with recent FSS methods on LVIS-92^i to further evaluate its effectiveness.
- **G Complementarity of Target-Aware Tokens and Spatial Prior**
  We investigate the complementary contributions of MLLM-derived target-aware tokens and support-guided spatial prior through ablation experiments.
- **H Limitations and Future Work of MA-FSS**
  We discuss the limitations of MA-FSS and directions for future improvement.
- **I Ethical Statement and Dataset Specification**
  We discuss responsible use, dataset maintenance, and potential limitations of VastFSS.

A CATEGORY INVENTORY AND DATA SOURCES

VastFSS comprises 2,400 categories, including 2,057 categories derived from V3Det (Wang et al., 2023) and 343 categories from datasets covering specialized applications. To support transparency and reproducibility, we provide the complete category inventory and document the corresponding data sources.

A.1 CATEGORIES FROM SPECIALIZED APPLICATIONS

Table 8 lists the data sources and associated categories for the 343 categories collected beyond V3Det. For ease of reference, we group these sources by application area, such as medical imaging, remote sensing, and solar physics.

Table 8: Data sources and associated categories from specialized applications in VastFSS.

Application Area	Data Source	Categories
Medical	HVSMR-2.0 (Pace et al., 2024)	left atrium; right atrium; thoracic aorta; inferior vena cava; left ventricle; right ventricle; pulmonary artery; superior vena cava
Medical	Intestinal segmentation dataset (Zhong et al., 2025)	appendix; ascending colon; cecum; descending colon, duodenum; rectum; sigmoid colon; small intestine; stomach; transverse colon

Application Area	Data Source	Categories
Medical	CholecInstanceSeg (Albabi et al., 2025)	bipolar; grasper; irrigator
Medical	OIMHS (Ye et al., 2023)	choroid; intraretinal cysts; macular hole
Remote Sensing	iSAID (Zamir et al., 2019)	bridge; ground track field; harbor; helicopter; large vehicle; roundabout; small vehicle; soccer ball field; swimming pool; baseball diamond; basketball court; plane; ship; storage tank; tennis court
Medical	FeTA (Payette et al., 2021)	cerebellum cortical gray matter; intracranial volume; lateral ventricles; subcortical gray matter; white matter
Medical	NasalSeg (Zhang et al., 2024b)	left maxillary sinus; left nasal cavity; nasal pharynx; right maxillary sinus; right nasal cavity
Medical	Dresden Surgical Anatomy Dataset (Carstens et al., 2023)	abdominal wall; colon; inferior mesenteric artery; intestinal veins; surgical liver; surgical pancreas; surgical small intestine; surgical spleen; surgical stomach; ureter; vesicular glands
Medical	FIVES (Jin et al., 2022)	fundus vessel
Biology	Pannuke (Gamper et al., 2019)	connective soft tissue cells; dead cells; epithelial cells; inflammatory cells; neoplastic cells
Industrial	DsCGF-3 (Lv et al., 2025)	foreign object; coal
Medical	IDRiD (Porwal et al., 2018)	haemorrhages; hard exudates; microaneurysms; optic disc; soft exudates
Medical	CT-ORG (Rister et al., 2020)	bladder; bone; kidneys
Medical	DIOME (Liu et al., 2024a)	tympanic membrane; cochlear promontory; incus; malleus; stapes
Remote Sensing	38-Cloud (Mohajerani & Saeedi, 2019)	cloud
Agricultural Remote Sensing	Agriculture-Vision (Chiu et al., 2020)	double plant; drydown; endrow; nutrient deficiency; planter skip; storm damage; waterway; weed cluster
Remote Sensing & Environment	CWLD (Lin et al., 2024)	building and facilities area; vacant landfill; waste accumulation area
Biology	PW-BALFC (Shi et al., 2025)	ciliated columnar epithelial; eosinophil cells; erythrocyte; lymphocyte; macrophage; neutrophil
Materials	SorpVision (Kabir et al., 2025)	water absorption region
Environment	CDW-Seg (Sirimewan et al., 2025)	waste fabric; timber; waste cardboard; concrete; fill dirt; hard plastic; plasterboard; soft plastic; steel

Application Area	Data Source	Categories
Marine Remote Sensing	Sea-Ice Segmentation Dataset (Kang et al., 2022)	sea ice
Biology	MoNuSeg (Kumar et al., 2017)	nuclei
Biology	Fluorescent Neuronal Cells dataset (Morelli et al., 2021)	fluorescent neuronal cells
Aerospace Science	NCSTP (Liu et al., 2025b)	antenna; observation payload; solar panel; satellite body
Medical	UAAL (Hao et al., 2025)	glottic aperture; left nostril; left vocal fold; right nostril; right vocal fold
Remote Sensing & Geological Hazard Monitoring	Landslide4Sense Dataset (Ghorbanzadeh et al., 2022)	landslide
Medical	Endoscapes (Mascagni et al., 2025)	cystic artery; cystic duct; cystic plate; surgical gallbladder; surgical tool
Medical	AIDK (Sun et al., 2024b)	keratitis lesion
Under Water	Seaclear Marine Debris Dataset (Duras et al., 2024)	fish; shells; sponge; starfish; other animal; urchin; plastic bag; glass bottle; plastic bottle; wood branch; clay brick; metal cable; metal can; paper cardboard; middle-size metal container; plastic container; ceramic cup; plastic cup; wood furniture; glass jar; plastic lid; plastic net; plastic pipe; plant; fiber rope; plastic rope; rov cable; rov tortuga; plastic sanitaries; paper snack wrapper; plastic snack wrapper; plastic tarp; rubber tire; cement tube; metal wreckage
Medical	BTCV (Landman et al., 2015)	IVC; aorta; liver; pancreas; spleen; veins
Medical	ISIC 2018 (Codella et al., 2019)	skin lesion
Medical	MTDDH (Qi et al., 2025)	ilium; obturator foramen
Medical	Children's Dental Panoramic Radiographs Dataset (Zhang et al., 2023)	adult tooth
Remote Sensing	LoveDA (Wang et al., 2021)	agriculture area; barren; building; road; water area
Medical	BTXRD (Yao et al., 2025)	osteosarcoma

Application Area	Data Source	Categories
Biomimetic Vision	COD10K (Fan et al., 2022a)	camouflaged ant; camouflaged cat; camouflaged caterpillar; camouflaged cicada; camouflaged deer; camouflaged duck; camouflaged gecko; camouflaged leafy sea dragon; camouflaged leopard; camouflaged mantis; camouflaged pipefish; camouflaged sea horse; camouflaged spider; camouflaged stick insect
Medical	MNI-HISUB25 (Kulaga-Yoskovitz et al., 2015)	cornu ammonis 1, 2, and 3 (CA1-3); cornu ammonis 4 and dentate gyrus (CA4-DG); subiculum
Remote Sensing & Disaster Management	Hurricane-Induced Debris Segmentation Dataset (Amini et al., 2025)	high-density debris; low-density debris
Civil Engineering	Indian Footpath Damage Segmentation Dataset (Chakurkar & Vora, 2025)	footpath damage
Medical	PSFHS (Chen et al., 2024)	fetal head; pubic symphysis
Solar Physics	SWEFil (Reche & Cid, 2024)	active region filaments (ARF); intermediate eegion filaments (IRF); quiescent region filaments (QRF); sunspot
Medical	TOM500 (Zhang et al., 2025)	inferior rectus; lacrimal gland; lateral rectus; medial rectus; optic nerve; orbital fat; superior oblique; superior rectus
Geological Engineering	GeoCrack (Yaqoob et al., 2024)	fracture edge
Medical	Germinoma MRI/CT and Radiomic Dataset (Huang et al., 2025)	germinoma
Medical	FracAtlas (Abedeen et al., 2023)	bone fracture
Medical	Kvasir-seg (Jha et al., 2020)	polyp
Industry	KolektorSDD (Tabernik et al., 2020)	surface defect
Environmental Remote Sensing	Oil Spill Drone Dataset (De Kerf et al., 2024)	oil spills
Medical	DDTI (Pedraza et al., 2015)	thyroid nodule
Cultural Heritage	MuralDH (Xu et al., 2024b)	mural damage

Application Area	Data Source	Categories
Medical	NPC MRI Dataset (Li et al., 2025)	nasopharyngeal carcinoma
Medical	CDD-CESM dataset (Khaled et al., 2022)	abnormal mammography
Medical	MSLesSeg (Guarnera et al., 2025)	multiple sclerosis lesion
Medical	BUS-UCLM (Vallez et al., 2025)	breast ultrasound lesion
Medical	BUSI (Al-Dhabyani et al., 2020)	malignant breast cancer
Remote Sensing & Energy Systems	Distributed solar photovoltaic array location and extent dataset (Bradbury et al., 2016)	solar photovoltaic (PV) arrays
Agriculture	PlantVillage (Mohanty et al., 2016)	apple scab; apple black rot; cedar apple rust; cherry powdery mildew; corn northern leaf blight; grape black rot; grape Esca; grape isariopsis leaf spot; orange citrus greening; peach bacterial spot; pepper bacterial spot; potato early blight; potato late blight; tomato bacterial spot; tomato early blight; tomato late blight; tomato leaf mold; tomato septoria leaf spot; tomato target spot; tomato yellow leaf curl virus; tomato mosaic virus
Human Parsing	LIP Dataset (Gong et al., 2017)	human beard; human left ear; human left eye; human lips; human nose; human right ear; human right eye

Application Area	Data Source	Categories
Under Water	FathomNet (Katija et al., 2022)	liponema brevicorne; caridea; squat lobster; asbestopluma; poeobius meseres; lycodes cortezianus; shrimp; plesionika; peniagone gracilis; galatheoidea; isosicyonis; asbestopluma monticola; actiniaria; munidopsis; black coral; enteroctopus; chionoecetes tanneri; asteroidea; squid; siphonophorae; hexactinellida; crab; chorilia longipes; moloha faxoni; funiculina; paralomis multispina; umbellula; modeeria rotunda; octopus; psolus squamatus; acanthoptilum; stony coral; colossendeis; amphipoda; boreomysis californica; glyphocrangon vicaria; pleuronectiformes; peniagone vitrea; hormathiidae; farrea; brisingida; pannychia; sebastolobus; octocorallia; sea sponge; lycodes diapterus; ophiacanthidae; synallactidae; crossaster borealis; nanomia bijuga; synallactes; sebastolobus altivelis; pennatulacea; boreomysis; hymenodora; pandalus amplus; bathochordaeus; psamminidae; stauro-calyptus; physonectae; apostichopus leukothele; gersemia juliepackardae; elpidia; chrysogorgia; caecosagitta macrocephala; munnopsidae; strongylocentrotus fragilis; metridium farcimen; ceriantharia; polynoidae; corbitellinae; ophiuroidea; comatulida; heterocarpus; gastropoda; kophobelemnon; aegina; ptychogena lactea; lycenchelys; eusergestes similis; sebastes cortezi; porifera; acanella weberi; mediaster aequalis; heteropolypus ritteri; pandalus platyceros; munida; bathyraja kincaidii; soft coral

A.2 Categories Derived from V3Det

The 2,057 categories derived from V3Det (Wang et al., 2023) are listed below in alphabetical order. The number in parentheses indicates the number of categories beginning with the corresponding letter.

A (81 classes)

abacus; *absinth*; *abutilon fruticosum*; *abutilon megapotamicum*; *acaena novae-zelandiae*; *acarine*; *accessory; acerola*; *achimenes; achyrachaena mollis*; *acmispon americanus*; *aconite*; *aconitum columbianum*; *acorn*; *aegonychon purpurocaeruleum*; *african daisy*; *african violet*; *agastache foeniculum*; *air breathing apparatus*; *air conditioner*; *air duct*; *ajuga chamaepitys*; *alderfly*; *alfalfa*; *alice band*; *allium schoenoprasum*; *aluminum alloy ladder*; *aluminum plastic tube*; *amaranth*; *amaranths*; *amelanchier utahensis*; *american dog violet*; *american football*; *american twinflower*; *ammannia coccinea*; *ammunition; anchusa arvensis*; *ancient qin*; *ancient zheng*; *anemopsis californica*; *angel's trumpet*; *anise*; *anisomorpha buprestoides*; *ankle support*; *annelid*; *ant*; *ant thrush*; *antenna*; *antennaria rosea*; *anthurium*; *anti-theft buckle hoist*; *antique*; *apocynum androsaemifolium*; *apple*; *apple of peru*; *applique*; *apricot*; *apron*; *aquarium*; *aquilegia chrysantha*; *arabic tea*; *arctotheca calendula*; *argemone albiflora*; *argiope bruennichi*; *arm guard*; *armband; armchair*; *armed response vehicle*; *armor*; *arnica; ashtray*; *aster*; *astragalus canadensis*; *astragalus danicus*; *astrophyton muricatum*; *asystasia gangetica*; *ataulfo*; *aurelia aurita*; *autumn sneezeweed*; *avifauna*; *awl*.

B (150 classes)

baby & toddler shoe; *baby bottle*; *baby float*; *back brace*; *backboard*; *badminton*; *bag*; *bag, wallet*; *bahiopsis parishii*; *baileya multiradiata*; *bald-faced hornet*; *ball*; *ball lock*; *ballpoint pen*; *balsa raft*; *balsamroot*; *bamboo*; *bamboo braiding*; *bamboo steamer*; *banana peel*; *bandeau*; *bandy stick*; *bar soap*; *barbell rack*; *barber's pole*; *barn*; *base layer*; *baseball*; *baseball bat*; *baseball uniform*; *basin*; *basket star*; *basketball*; *basketball moves*; *bath heater*; *bath towel*; *battering ram*; *battery*; *batting glove*; *battle ropes*; *beach towel*; *bead*; *bean*; *bear grass*; *bearing hinge*; *bebbia juncea*; *bed*; *bed pillow*; *bee*; *bee balm*; *beet green*; *bell*; *bell-clapping*; *bellardia trixago*; *belt*; *bench (weight training)*; *bench press*; *bend bar*; *berry*; *berteroa incana*; *betonica officinalis*; *beverage*; *bianqing*; *bicycle bell*; *bicycle handlebar*; *bicycle seatpost*; *bicycle wheel rim*; *bicycle-built-for-two*; *bidens aristosa*; *bidens cernua*; *bidens pilosa*; *big brown bat*; *big Chinese cabbage*; *bilberry*; *billboard*; *billiards*; *billy buttons*; *bindweed*; *bingsu*; *bird of paradise*; *bird's foot trefoil*; *bird's nest*; *bird's-foot violet*; *birdhouse*; *bistorta bistortoides*; *bistorta officinalis*; *bittercress*; *bitters*; *black belt (martial arts)*; *black hole*; *black medick*; *black mulberry*; *black pudding*; *black-footed albatross*; *blackstonia perfoliata*; *blanched shrimp*; *blanket*; *blindfold*; *bloodworm*; *blowgun*; *blue columbine*; *blueberry yogurt*; *bluetooth earphone*; *blush wine*; *board*; *boat trailer*; *boater*; *bobbin*; *bog rosemary*; *bok choy*; *bone conduction headphones*; *bookmark*; *boot*; *bottle*; *bottle opener*; *bottled water*; *bow*; *bowl*; *bowling*; *box truck*; *boxing mouthguard*; *bracelet*; *bramble fruit*; *brassica rapa*; *brassiere*; *breakaway rim*; *bream*; *bridge*; *briefcase*; *bristly locust*; *brittle star*; *broom*; *brown pelican*; *buccinum undatum*; *bucket*; *buckle*; *buckthorn berry*; *building*; *bullet*; *bulletproof clothing*; *bumper car*; *bung*; *bus*; *bush violet*; *business card*; *butter*; *butter paper*; *butterwort*; *bánh tét*; *bánh xèo*.

C (238 classes)

cabana; *cabinet*; *cabinet air conditioner*; *cafe noir*; *cage*; *calamansi*; *calendula*; *calendula arvensis*; *california bluebell*; *california four o'clock*; *calliandra*; *calliandra eriophylla*; *callirhoe involucrata*; *calycanthus floridus*; *calyptocarpus vialis*; *calystegia macrostegia*; *camail*; *camcorder*; *camellia*; *camomile*; *campanula glomerata*; *campanula prenanthoides*; *campanula trachelium*; *camper*; *camper trailer*; *canada anemone*; *canada dry*; *candied apple*; *candle*; *candlesnuffer*; *candy*; *candy cane*; *canopy bed*; *cap*; *cape gooseberry*; *cape jasmine*; *capri pants*; *carabiner*; *caranday*; *carbon paper*; *cardamine flexuosa*; *cardboard*; *cardigan welsh corgi*; *cargo trailer*; *carolina spring beauty*; *carpenteria*; *carpobrotus chilensis*; *carton*; *cash machine*; *cashmere sweater*; *cassava*; *castilleja affinis*; *castilleja applegatei*; *castor bean*; *casual pants*; *catheter*; *cattail*; *cattley guava*; *cauliflower*; *ceanothus herbaceus*; *ceiba pentandra*; *ceiling fan*; *celandine*; *celery*; *celestial body*; *cell phone*; *celtuce*; *cenotaph*; *cerastium glomeratum*; *chain lock*; *chair*; *chairlift*; *chalk*; *chaparral mallow*; *chaparral pea*; *charger*; *checkerbloom*; *cheerleading uniform*; *cheese*; *cherimoya*; *cherry*; *chess piece*; *chess set*; *chess table*; *chest protector*; *chewing gum*; *chicken coop*; *chiffon shirt*; *child and mother knife*; *chinese forget-me-not*; *chinese lantern plant*; *chinese magnolia*; *chinese peony*; *choc ice*; *chocolate*; *chokeberry*; *chokecherry*; *chopsticks*; *chrysanthemum*; *chrysoberyl*; *chuck wagon*; *chuckwagon*; *chuquiraga jussieui*; *chāhan*; *cidomo*; *cigarette*; *cigarette butt*; *cinquefoil*; *citrange*; *citrus*; *citrus trifoliata*; *clam*; *clarkia amoena*; *clarkia rhomboidea*; *clay pot*; *cleaning pad*; *cleaning implement*; *clematis alpina*; *clematis cultivar*; *clematis lasiantha*; *cleome rutidosperma*; *clingfish*; *clipboard*; *cloth covering*; *clothes tree*; *clothing in tibet*; *clupeid fish*; *clusia*; *coat hanger*; *coatrack*; *cobra*; *cobweb*; *coca-cola*; *cochlospermum vitifolium*; *cockhorse*; *cockle*; *coco plum*; *cocoa powder*; *coconut*; *coffee*; *coffee liqueur*; *coffee maker*; *coinage*; *collar*; *colt's foot*; *columbiform bird*; *comb*; *comb-footed spider*; *combine with bowl*; *comforter*; *commelina communis*; *common foxglove*; *Common periwinkle*; *common wood sorrel*; *compact car*; *compass plant*; *component*; *compression bandage*; *computer desk*; *conch*; *concrete mixer*; *condom*; *conduit*; *confederate rose mallow*; *conference table*; *conicosia pugioniformis*; *consolida regalis*; *contact grill*; *container*; *convolvulus*; *convolvulus arvensis*; *cordial*; *coreid bug*; *coreopsis*; *corethrogyne filaginifolia*; *cork*; *corkboard*; *cormous plant*; *corn chip*; *corn cockle*; *corn marigold*; *Corner hinge*; *cornus canadensis*; *correction fluid*; *corsage*; *corydalis flavula*; *corylus avellana*; *coryphantha macromeris*; *cosmetic container*; *cosmos*; *cosmos bipinnatus*; *cottage tent*; *couplet*; *coverall*; *cowboy hat*; *cowrie*; *CPU*; *crab cocktail*; *cranberry*; *crane*; *cranesbill*; *cream liqueur*; *cream yogurt*; *creamcups*; *creme de cacao*; *crepidula fornicata*; *crepis capillaris*; *cricket*; *criollo*; *crop top*; *croquet*; *croquet mallet*; *cross*; *cross of sacrifice*; *crossbow*; *crosswalk*; *crosswalk sign*; *croton glandulosus*; *crown daisy*; *crown of thorns*; *crustacean*; *crutch*; *crème de cassis*; *cuba libre*; *Cucurbita*; *cue stick*; *cup*; *curler*; *currant*; *curtain*; *curtain rod*; *cutting implement*; *cymbal*; *cyperus echinatus*; *cyperus eragrostis*; *cypress vine*; *cyprinid*; *cypripedium candidum*.

D (91 classes)

dairy product; *daisy*; *dalea purpurea*; *damselfly*; *dashiki*; *data line*; *date plum*; *datura inoxia*; *datura stramonium*; *deck house*; *decoration*; *delivery truck*; *delphinium nudicaule*; *delphinium parishii*; *dendraster excentricus*; *dendromecon rigida*; *dental appliance*; *dermasterias imbricata*; *desert rose*; *desk*; *detector*; *dianthus barbatus*; *dianthus carthusianorum*; *dianthus superbus*; *dice cup*; *dichelostemma capitatum*; *die*; *diervilla lonicera*; *digitalis grandiflora*; *dim sum*; *dinner table*; *diospyros virginiana*; *diplacus aurantiacus*; *diplacus bigelovii*; *diplacus longiflorus*; *diplacus puniceus*; *diploma*; *direct vent fireplace*; *discharge pipe*; *discus*; *disguise*; *dish*; *dish rack*; *dishwasher*; *display device*; *diving helmet*; *diving suit*; *doellingeria umbellata*; *dog collar*; *dog fashion*; *doily*; *doll*; *doll clothes*; *dolls' pram*; *dolsot*; *door*; *door jamb*; *doorframe*; *doorknob*; *doorplate*; *dormer*; *double bar towel rack*; *double cup holder*; *dough*; *down vest*; *dragonfly*; *dragoon helmet*; *drawer*; *drawer rail*; *dray*; *dreidel*; *dress shirts*; *dresser*; *dressing table*; *drift trike*; *drill*; *drinking water*; *drop cloth*; *drum stick*; *drumhead*; *dryas octopetala*; *dublin dr pepper*; *duck king*; *duck liver*; *dumbbell*; *dumbbell stand*; *dumpster*; *duranta erecta*; *durian*; *dustpan*; *dutchman's-pipe*.

E (70 classes)

echinacea purpurea; *echinocereus coccineus*; *echinocereus enneacanthus*; *echinocereus pectinatus*; *echinocereus reichenbachii*; *echinocereus santaritensis*; *echinocereus triglochidiatus*; *echinometra mathaei*; *echinops ritro*; *edible snail*; *edible_nut*; *egg as food*; *egg yolk*; *elastic device*; *elbow pad*; *elderberry*; *electric fan*; *electric range*; *electric trike*; *electrically tricycle*; *electronic component*; *electronic device*; *emblem*; *emilia fosbergii*; *emilia sonchifolia*; *energy saving lamp*; *english plantain*; *ensemble*; *entelea arborescens*; *entrance*; *epaulette*; *epergne*; *epidendrum radicans*; *epigaea repens*; *epilobium brachycarpum*; *epilobium ciliatum*; *epilobium montanum*; *epipactis atrorubens*; *epiphytic cactus*; *episcia*; *eriastrum eremicum*; *erica abietina*; *erigeron acris*; *eriocapitella hupehensis*; *eriogonum latifolium*; *eriophorum vaginatum*; *erodium botrys*; *erodium cicutarium*; *eryngium planum*; *erythranthe alsinoides*; *erythranthe guttata*; *erythranthe lewisii*; *erythranthe moschata*; *eschscholzia caespitosa*; *eschscholzia parishii*; *escort carrier*; *espresso*; *eucrypta chrysanthemifolia*; *euonymus verrucosus*; *euphorbia albomarginata*; *european lobster*; *european sanicle*; *euryops pectinatus*; *evening primrose*; *exercise bike*; *exhaust fan*; *exocarpos cupressiformis*; *expansion bolt*; *external screen*; *eyeglass lens*.

F (79 classes)

fabric; *face powder*; *false foxglove*; *false mallow*; *fanta*; *faucet*; *feather boa*; *fedora*; *fennel*; *ferocactus emoryi*; *field scabious*; *field speedwell*; *field tent*; *fiesta flower*; *fig*; *figurine*; *file*; *files rack*; *fingerless glove*; *fingerpost*; *fire emergency lights*; *fire extinguisher*; *fire mask*; *fire tongs*; *fireplug*; *fish*; *fishing pliers*; *fishing reel*; *fishing rod*; *fishing gear*; *fitness ring*; *fitness stick*; *flag*; *flagpole*; *flat bench*; *flat ladder*; *flipper*; *floor cover*; *floor fan*; *floor lamp*; *flower*; *flowering maple*; *flowering tobacco*; *fluorescent*; *flute*; *foam shaft*; *folding chair*; *folding napkins*; *food steamer*; *food truck*; *footwear*; *forehead mirror*; *fork*; *forked weapon*; *fountain*; *four legged table*; *four loko*; *four o'clock*; *foxtail*; *framework*; *frasera caroliniensis*; *freetail*; *freezer*; *french door*; *french honeysuckle*; *fried egg*; *fringe*; *fringed gentian*; *fringed loosestrife*; *frozen yogurt*; *frozen dessert*; *fruit salad*; *fruit yogurts*; *fuchsia boliviana*; *full-face helmet*; *fumaria capreolata*; *fur collar*; *fur hat*; *fustanella*

G (97 classes)

galanthus nivalis; *galeopsis bifida*; *galia*; *game controller*; *garbage*; *garden hose*; *garden tool*; *garlic*; *garment steamer*; *gas oven*; *gateleg table*; *gaultheria depressa*; *gauntlet*; *gear*; *gearshift*; *gebakkraam*; *general motors*; *generator*; *genip*; *gentianella*; *geoduck*; *geraea canescens*; *geranium*; *geranium dissectum*; *geranium palustre*; *geranium phaeum*; *geranium purpureum*; *geranium pusillum*; *geranium pyrenaicum*; *geranium robertianum*; *geranium sanguineum*; *geranium sylvaticum*; *gerardia*; *geum aleppicum*; *geum rossii*; *geum triflorum*; *ghost bike*; *gift wrap*; *gilia tricolor*; *ginger*; *ginger beer*; *ginger buds*; *glass*; *glass goblet*; *glass hinge*; *eyeglasses*; *glasses*; *glasses case*; *globe-flower*; *glove*; *gloxinia*; *Go*; *go-kart*; *goalmouth*; *goatsfoot*; *godiva liqueur*; *goji*; *goji tea*; *golden barrel cactus*; *golden clematis*; *golf cart*; *golf club*; *gomphocarpus physocarpus*; *goodenia ovata*; *gorgonian*; *gourd mouth organ*; *grab bar*; *grain*; *grape juice*; *grapefruit juice*; *grapes*; *grapnel*; *grattachecca*; *gravy boat*; *great barracuda*; *greater burdock*; *greek valerian*; *greek vase*; *green*

bell pepper; *green onion*; *green pepper*; *greens*; *greeting card*; *griddle*; *grilled zucchini*; *grindelia stricta*; *gromwell*; *groundsheet*; *gryllus pennsylvanicus*; *guitar*; *gumdrop*; *gumweed*; *gun*; *gun barrel*; *gundam model*; *gypsum carving*; *gyūdon*.

H (96 classes)

hair accessory; *hair band*; *hair clip*; *halberd*; *half zip sweater*; *half-circle skirt*; *hammer*; *hand cream*; *hand truck*; *hand tool*; *handball*; *handcuff*; *handkerchief*; *handle lock*; *handwheel*; *hang flag*; *Hang lock*; *hang tag*; *hardware*; *hardy kiwi*; *hare wallaby*; *harebell*; *harem pants*; *harp*; *hat*; *hawkweed*; *hawthorn*; *hayloft*; *HDMI interface*; *headband, head covering, hair accessory*; *headboard*; *headdress*; *headphone ear pad*; *headscarf*; *hearing aid*; *hearth*; *heartseed*; *heater*; *height tester*; *helenium amarum*; *helenium bigelovii*; *helianthemum*; *helianthemum nummularium*; *helianthus angustifolius*; *helianthus divaricatus*; *helianthus mollis*; *helianthus occidentalis*; *heliotropium indicum*; *helmet*; *helminthotheca echioides*; *hemizonia congesta*; *hemp nettle*; *henbane*; *hepatica acutiloba*; *herb bennet*; *heterotheca subaxillaris*; *hex nut*; *hexagonal trampoline*; *high heels*; *highchair*; *hippeastrum*; *hippopotamus amphibius*; *hoary plantain*; *hockey*; *hockey stick*; *hoe*; *holocnemus pluchei*; *holster*; *home appliance*; *home brew*; *home plate*; *honey dipper*; *hong kong-style milk tea*; *hood*; *hoodie*; *hook*; *hooker's onion*; *hoop*; *horizontal bar*; *horse nettle*; *horseless carriage*; *horsemint*; *horseshoe*; *hot pepper*; *house*; *housing*; *houstonia longifolia*; *hula-hoop*; *hulusi*; *hurdle*; *hydrangea macrophylla*; *hydrangea quercifolia*; *hydrant*; *hymenocallis littoralis*; *hypochaeris glabra*; *hypodermic needle*.

I (32 classes)

ice cream; *ice cream cart*; *ice cream cone*; *iced coffee*; *icelandic turf house*; *impala lily*; *impatiens balfourii*; *impatiens glandulifera*; *impatiens parviflora*; *impatiens walleriana*; *indian mallow*; *indigofera miniata*; *inflatable armbands*; *injection*; *inner tube*; *insect*; *instrumentality*; *inula*; *inula britannica*; *inula helenium*; *inula salicina*; *invitation*; *ipomoea batatas*; *ipomoea cordatotriloba*; *ipomoea indica*; *ipomoea triloba*; *ipomopsis longiflora*; *iridaceous plant*; *iron artwork*; *irrigation sprinkler*; *ixeris chinensis*; *ixora coccinea*.

J (31 classes)

jaboticaba; *jackfruit*; *jamaica rum*; *japanese chestnut*; *japanese persimmon*; *japanese pink*; *japanese quince*; *japanese snowbell*; *japchae*; *jasione montana*; *jasmine*; *jaumea carnosa*; *jaunting car*; *javelin*; *denim shorts*; *jeans*; *jeffersonia diphylla*; *jello*; *jelly babies*; *jersey*; *jerusalem artichoke*; *jinrikisha*; *job's tears*; *jointed charlock*; *joss paper*; *joystick*; *jug*; *juice*; *jujube*; *jump rope*; *juniper berry*

K (33 classes)

kabayaki; *kaffiyeh*; *kallstroemia grandiflora*; *kalmia microphylla*; *kang-ku*; *kassler*; *kayak paddle*; *keg*; *kerria japonica*; *kettle*; *kettlebell*; *key*; *keychain*; *kilt*; *Kino bamboo drum*; *kitchen sink*; *kitchen appliance*; *kitchen utensil*; *kitchenware*; *kite*; *kiwi fruit*; *knautia arvensis*; *knee pad*; *knife*; *knit shirt*; *knitwear*; *knocker*; *kohleria*; *kohlrabi*; *koinobori*; *kosteletzkya pentacarpos*; *krameria lanceolata*; *kumquat*

L (91 classes)

lacrosse; *lacrosse ball*; *lactuca floridana*; *lagurus ovatus*; *lake dwelling*; *lamp*; *lapel*; *large periwinkle*; *large-flowered cactus*; *larix occidentalis*; *lasthenia gracilis*; *laundry detergent*; *layia glandulosa*; *lcd projector*; *leaf-footed bug*; *leafless tree*; *leather*; *leather clothing*; *Leather-covered notebook*; *lectern*; *leek*; *leg warmer*; *lego*; *lekor*; *lemon*; *lemon liqueur*; *lemon tea*; *lemonade*; *lemon-soda*; *lens*; *leonotis leonurus*; *leontopodium nivale*; *leopard plant*; *leopard's-bane*; *leptosiphon nuttallii*; *leptosiphon parviflorus*; *leptospermum scoparium*; *leucojum vernum*; *leycesteria formosa*; *lianxiang stick*; *liatris cylindracea*; *liatris spicata*; *life jacket*; *life saver*; *Lifting drying rack*; *Light bulb*; *lighter*; *lighthouse*; *lighting fixture*; *liliaceous plant*; *limeade*; *limodorum abortivum*; *linanthus pungens*; *lingonberry*; *linnaea borealis*; *linum bienne*; *linum lewisii*; *lipstick plant*; *liquor*;

litchi; *lithophyte*; *lithospermum canescens*; *little golden zinnia*; *livingstone daisy*; *lobster-claws*; *lock*; *locking pliers*; *long sleeve T-shirt*; *long spear*; *longbow*; *loofah*; *lopezia racemosa*; *loquat*; *lotion*; *lotus root*; *loudspeaker*; *louver*; *love-in-a-mist*; *ludwigia octovalvis*; *ludwigia peploides*; *ludwigia peruviana*; *lupinus arcticus*; *lupinus nanus*; *lupinus polyphyllus*; *lupinus succulentus*; *lute*; *lycium barbarum*; *lysimachia borealis*; *lysimachia maritima*; *lysimachia nemorum*; *lysimachia thyrsiflora*

M (101 classes)

M16A4; *macadamia nut*; *magazine*; *maglev crafts*; *magnetic tape*; *magnifying glass*; *mail slot*; *mallee*; *malva arborea*; *malva moschata*; *malva neglecta*; *malvastrum coromandelianum*; *malvaviscus arboreus*; *mammillaria heyderi*; *manduca sexta*; *mangifera indica*; *mangosteen*; *mannequin*; *maraschino*; *marker*; *marrow*; *mars*; *marsh st-john's wort*; *masala chai*; *mask*; *masonry oven*; *mat*; *matilija poppy*; *matryoshka doll*; *mausoleum*; *maxiskirt*; *mayweed*; *mbira*; *meadow saxifrage*; *measuring instrument*; *meat*; *mecardonia procumbens*; *mechanical device*; *medal*; *medical instrument*; *medlar*; *medusa*; *megaphone*; *megger*; *melampyrum arvense*; *melampyrum nemorosum*; *melastoma malabathricum*; *melochia pyramidata*; *melochia tomentosa*; *melon and jamón*; *mentha aquatica*; *mentzelia involucrata*; *menu*; *Mercury*; *meteor hammer*; *micranthes californica*; *micrathena gracilis*; *microphone*; *microsteris gracilis*; *microwave*; *military uniform*; *milkshake*; *mimulus ringens*; *mincer*; *mineral water*; *minibar*; *mirinda*; *mirror nail*; *missouri primrose*; *mixing console*; *mobile device*; *moccasin*; *model*; *module*; *moisture-proof box*; *molluscs*; *mombin*; *monarda*; *monardella odoratissima*; *moneywort*; *monitor*; *monster truck*; *montia parvifolia*; *moppingfloor*; *morin khur*; *morning glory (plant genus)*; *mountain ebony*; *mountain everlasting*; *mouse-ear chickweed*; *mud dauber*; *muff*; *mug*; *muilla maritima*; *mulled wine*; *mullein pink*; *murraya paniculata*; *musical keyboard*; *musical saw*; *mustang mint*; *myosotis arvensis*; *myrtle*

N (40 classes)

nail; *nail polish*; *nama demissa*; *nama hispida*; *napa cabbage*; *natal plum*; *neck brace*; *necklace*; *neckline*; *neckpiece*; *necktie*; *nectar*; *needle*; *nematode*; *nemophila heterophylla*; *nemophila maculata*; *nemophila menziesii*; *nerf*; *nest*; *newspaper*; *newsstand*; *newton's cradle*; *nidus*; *night-blooming cactus*; *nightstand*; *nightwear*; *nine men's morris*; *nissen hut*; *noodles*; *northern jacob's ladder*; *norway lobster*; *notepad*; *notepaper*; *nothoscordum bivalve*; *nucella lamellosa*; *nunchaku*; *nuphar polysepala*; *nut*; *nuttallanthus canadensis*; *nymphoides indica*.

O (34 classes)

oeil de boeuf; *oenothera berlandieri*; *oenothera biennis*; *oenothera cespitosa*; *oenothera elata*; *oenothera laciniata*; *oenothera rosea*; *oenothera speciosa*; *okra*; *olsynium douglasii*; *omnidirectional wheel*; *one-legged table*; *onion*; *opuntia fragilis*; *opuntia stricta*; *orange bitters*; *orange juice*; *orange sneezeweed*; *orchid*; *orchis purpurea*; *orobanche minor*; *osage orange*; *osteospermum moniliferum*; *outdoor grill*; *outerwear*; *oven*; *overgarment*; *oxalis articulata*; *oxalis corniculata*; *oxalis latifolia*; *oxalis montana*; *oxalis stricta*; *oxeye*; *oxeye daisy*

P (186 classes)

pachystachys lutea; *packera glabella*; *packera tampicana*; *pad*; *paeonia californica*; *paigu*; *palafoxia callosa*; *pan*; *pancratium maritimum*; *pandeiro*; *panel van*; *pansy*; *papaver dubium*; *papaver heterophyllum*; *papaver rhoeas*; *papaw*; *papaya*; *paper*; *paper bag*; *paper bowl*; *paper box*; *paper cup*; *paper cutter*; *paper plate*; *paper shelf*; *paper takeout box*; *paper towel*; *parallelbars*; *parasol mushroom*; *paratransit*; *partytent*; *passenger van*; *passerby*; *passiflora tripartita*; *passion fruit*; *path roller*; *pay-phone*; *PD interface*; *peace pole*; *Ped-o-Pull*; *pedelec*; *pedestal table*; *pedicab*; *pedicularis densiflora*; *peeler*; *pen holder*; *pencil*; *pencil sharpener*; *pennyroyal*; *pentaglottis sempervirens*; *peony*; *pepsi*; *percussion instrument*; *peritoma serrulata*; *perityle emoryi*; *periwinkle*; *persian violet*; *persicaria amphibia*; *persicaria lapathifolia*; *persicaria maculosa*; *pestle*; *pesäpallo*; *petunia*; *phacelia bipinnatifida*; *phacelia grandiflora*; *phacelia parryi*; *phacelia purshii*; *phlox divaricata*; *phlox drummondii*; *phlox hoodii*; *phoenicopterus roseus*; *pholistoma membranaceum*; *photographic_equipment*; *phulkari*; *phyla lanceolata*; *phyla nodiflora*; *phytolacca icosandra*; *pickerelweed*; *piece of cloth*; *pillbox*; *pillow*; *pimm's*; *pimpernel*; *pin*; *pineapple juice*; *ping pong net*;

pingpong; *pinguicula vulgaris*; *pink evening primrose*; *pinkladies*; *pinwheel*; *pipe tobacco*; *pipefitting*; *piranha*; *pisaster giganteus*; *pitcher plant*; *pityopsis falcata*; *plane*; *plastic*; *plastic art*; *plasticine*; *playing card*; *pliers*; *plumbago zeylanica*; *plumbing fitting*; *plus fours*; *poacher*; *pocket*; *pogonia ophioglossoides*; *police van*; *polo mallet*; *polo shirt*; *polygala lutea*; *polygala nana*; *polygala rugelii*; *polypremum procumbens*; *pomelo*; *pontederia crassipes*; *pony cart*; *poodle skirt*; *popcorn machine*; *popper*; *poppy*; *poppy mallow*; *popsicle*; *porkpie*; *porringer*; *portcullis*; *porthole*; *portulaca*; *portulaca pilosa*; *sticky note pad*; *post-it note*; *postcard*; *poster*; *pot*; *pot rack*; *cooked potato*; *potato*; *potentilla argentea*; *potentilla canadensis*; *potentilla erecta*; *potentilla gracilis*; *potentilla indica*; *potentilla norvegica*; *potentilla simplex*; *potherb*; *potholder*; *potty seat*; *Power Adapter*; *power plugs and sockets*; *prairie sunflower*; *prawn*; *prawn ball*; *pressure cooker*; *pressure washer*; *pressure_suit*; *primula pauciflora*; *primula vulgaris*; *projection screen*; *protea repens*; *protective_covering*; *protractor*; *proverb*; *prunus persica*; *pseudognaphalium californicum*; *psorothamnus schottii*; *pterostylis pedunculata*; *public_transport*; *puck*; *pudding corn*; *pull-up*; *pulley*; *pulmonaria obscura*; *pulsatilla alpina*; *pump well*; *pumpjack*; *pumpkin seed*; *punch bowl*; *punica granatum*; *puppet*; *purple cress*; *purse*; *pushing cart*; *pyracantha coccinea*; *pyramid bugle*.

Q (5 classes)

quandong; *qubba*; *quercus garryana*; *quercus rotundifolia*; *quiver*.

R (82 classes)

rabbit ears; *rabbit hutch*; *racket*; *radish*; *raft*; *rag doll*; *rain barrel*; *rainbow cactus*; *rake*; *ramune*; *ranch house*; *rapistrum rugosum*; *rattlesnake root*; *rawap*; *razor*; *rear-view mirror*; *record player*; *recumbent bicycle*; *red algae*; *red bean paste*; *red bell pepper*; *red kuri squash*; *redfish*; *refrigerator*; *refrigerator magnet*; *relief carving*; *remote*; *reptile*; *rescue whistle*; *rescue equipment*; *resistance training sticks*; *resistor*; *reusable shopping bag*; *reversi*; *revolver*; *revolving door*; *rhexia alifanus*; *rhexia mariana*; *rhexia virginica*; *rhodiola integrifolia*; *rhododendron canadense*; *rhododendron groenlandicum*; *rhododendron macrophyllum*; *ribbed frame*; *ribbon*; *ribbon candy*; *ribes lacustre*; *ribes nevadense*; *richardia grandiflora*; *rickshaw*; *riding helmet*; *riding mower*; *rifle scopes*; *rivet*; *rivet nut*; *robe*; *robinia neomexicana*; *rocket model*; *rocking chair*; *rodent*; *roller shoe*; *roller skates*; *rolling pin*; *roly-poly toy*; *root vegetable*; *rose*; *rotisserie*; *round table*; *round-headed leek*; *rubber band*; *rubber duck*; *rubber eraser*; *rubus allegheniensis*; *rubus ulmifolius*; *rudbeckia amplexicaulis*; *ruellia caroliniensis*; *ruellia humilis*; *ruellia strepens*; *ruffle*; *rugby ball*; *rugby goal*; *rutabaga*.

S (254 classes)

sacred lotus; *saddle*; *safety lock buckle*; *safflower*; *sagebrush buttercup*; *sainfoin*; *sake*; *salmon*; *salpiglossis*; *saltwater fish*; *salvia columbariae*; *salvia hispanica*; *salvia verticillata*; *samara*; *sambucus ebulus*; *samolus repens*; *sand lance*; *sand myrtle*; *sand spurry*; *sandals*; *sanguisorba canadensis*; *sanguisorba officinalis*; *sanicula arctopoides*; *sapayi*; *saskatoon*; *satellite*; *Saturn*; *sauce*; *saucepan*; *saxifraga bronchialis*; *saxifraga oppositifolia*; *scabiosa ochroleuca*; *scabious*; *scaevola plumieri*; *scallop shell*; *scarecrow*; *scarf*; *school desks*; *scissors*; *scorpion fly*; *screw*; *screw bean*; *screwdriver*; *scrunchie*; *scull*; *sculpture*; *scutellaria galericulata*; *scutellaria lateriflora*; *scythe*; *sea cucumber*; *sea otter*; *sea chub*; *seafood*; *seashell*; *securigera varia*; *sedan*; *seed*; *Seedling knife*; *seesaw*; *self-heal*; *self-tapping screw*; *selliera radicans*; *semitrailer*; *senecio flaccidus*; *senecio inaequidens*; *senna armata*; *senna covesii*; *sensitive plant*; *sentry box*; *sequin*; *sewing machine*; *shackle*; *shallot*; *shave ice*; *shaving brush*; *sheepskin boots*; *sheet pan*; *shelf*; *shell*; *shellfish*; *shieling hut*; *shin guard*; *shinai*; *shipping*; *shirt*; *shirt collar*; *shirt, blouse*; *shoe*; *Shoebox*; *shop*; *short skirt*; *short sleeve T-shirt*; *shorts*; *shoulder board*; *shoulder patch*; *shoulder sleeve insignia*; *shower*; *fixed showerhead*; *handheld showerhead*; *showy milkweed*; *shrubby st john's wort*; *shtreimel*; *shumai*; *sickle alfalfa*; *sida rhombifolia*; *sidalcea oregana*; *sidalcea sparsifolia*; *signage*; *signboard*; *signpost*; *silene antirrhina*; *silene flos-cuculi*; *silene laciniata*; *silene virginica*; *silk*; *silkworm*; *silphium albiflorum*; *silphium integrifolium*; *silphium perfoliatum*; *silphium terebinthinaceum*; *silversword*; *silverweed*; *single cup holder*; *single pole towel rack*; *single scull*; *singlet*; *sink*; *sippy cup*; *sisyrinchium bellum*; *sisyrinchium rosulatum*; *sit-ski*; *skateboard*; *skein*; *ski*; *ski boot*; *ski pole*; *skibob*; *skirt*; *sled*; *sleeve*; *slide*; *sliding door*; *sliding door track*; *slipper*; *smoke grenade*; *smoked scallop*; *smokestack*; *snapper*; *sneezeweed*; *snorkel*; *snowboard*; *soap*; *soap dish*; *sock*; *socket*; *socket wrench*; *sod house*;

soda water; *soft drink*; *softball bat*; *solanum douglasii*; *solanum lycopersicum*; *solanum umbelliferum*; *solanum xanti*; *soldering iron*; *sonchus arvensis*; *sorb*; *soup bowl*; *spade*; *spathoglottis plicata*; *speed sign*; *Speedometer*; *sphagneticola trilobata*; *spider*; *spiderwort*; *spigelia marilandica*; *spiraea splendens*; *spire*; *Spirometry*; *splitting maul*; *sponge (animal)*; *spoon*; *spoon lure*; *spork*; *sports ball*; *sports equipment*; *sportswear*; *spotted knapweed*; *spray paint*; *spring*; *spring beauty*; *spring cress*; *spring rolls*; *springform pan*; *sprout*; *spurge nettle*; *square table*; *stacked heel*; *stag beetle*; *stanleya pinnata*; *stapelia*; *staple gun*; *star apple*; *star-of-bethlehem*; *starting pistol*; *stationary bicycle*; *statuary*; *steak tartare*; *steering wheel*; *stellaria graminea*; *stellaria media*; *stellaria nemorum*; *stemware*; *stephanomeria pauciflora*; *stereo*; *stethoscope*; *stickpin*; *Stilts*; *stirrer*; *stirrup*; *stockings*; *stone cabin*; *stone carving*; *stop sign*; *storksbill*; *strandpulling*; *straw*; *strawberry geranium*; *street light*; *stretcher*; *string bells*; *stringed instrument*; *strophanthus*; *stun gun*; *summer squash*; *sun*; *sunchoke*; *sunglasses*; *surgeonfish*; *surgical mask*; *surinam cherry*; *suspenders*; *swatter*; *sweatband*; *sweater*; *sweet*; *sweet four o'clock*; *swing (seat)*; *swiss cheese*; *switch*; *sword*; *symphoricarpos rotundifolius*; *symphyotrichum divaricatum*; *symphyotrichum lanceolatum*

T (123 classes)

table knife; *table lamp*; *table tennis racket*; *table linen*; *tablet*; *tablet computer*; *tableware*; *tadarida brasiliensis*; *tagetes*; *tai chi knife*; *taipan*; *tall bellflower*; *tall oat grass*; *tamarind*; *tamarindus indica*; *tanacetum parthenium*; *tandem trike*; *tandoor*; *tap*; *tapestry*; *tarpaulin*; *tassel*; *taxi*; *tayberry*; *tea ball*; *tea maker*; *tea pot*; *teasel*; *teddy*; *teh tarik*; *telescope*; *telescopic sight*; *tennis net*; *tennis rack*; *tent*; *tequila*; *terrestrial globe*; *thalictrum thalictroides*; *thelesperma filifolium*; *thimble*; *thin cardigan*; *thistle*; *thrips*; *throw pillow*; *throwing knife*; *thunbergia alata*; *thunbergia grandiflora*; *thyme-leaved speedwell*; *thymophylla pentachaeta*; *tic-tac-toe*; *tick*; *tickseed sunflower*; *tie*; *tights*; *tights, stockings*; *timer*; *timer clock*; *timothy*; *tire*; *tissue*; *tithonia diversifolia*; *toaster*; *tobacco barn*; *toilet brush*; *toilet paper*; *toilet roll holder*; *toiletries*; *tollbooth*; *tomatillo*; *tomato*; *tomb*; *tomb of the unknown soldier*; *tonfa*; *tongs*; *toothbrush*; *toothpick*; *top, t-shirt, sweatshirt*; *topknot*; *topminnow*; *toque*; *torpedo tube*; *touchpad*; *towel*; *towel rack*; *toy*; *toy vehicle*; *tractor*; *tradescantia occidentalis*; *tradescantia ohiensis*; *traffic light*; *treadle machine*; *treadmill*; *tree*; *tree poppy*; *tree swallow*; *trekking pole*; *triangle*; *trichoptilium incisum*; *tricycle*; *tridax procumbens*; *trifoliate orange*; *trifolium aureum*; *trifolium fragiferum*; *trifolium montanum*; *trifolium subterraneum*; *trilobite*; *triodanis biflora*; *triteleia laxa*; *triumphal arch*; *trollius europaeus*; *trombone*; *trompo*; *tropaeolum majus*; *trophy*; *trowel*; *trumpetfish*; *trunks*; *tube*; *turkish coffee*; *turnera ulmifolia*; *turtleneck sweater*; *tv cabinet*; *tv set*.

U (15 classes)

u-lock; *ube ice cream*; *umbrella*; *umbrella tent*; *umpire chair*; *unadon*; *undergarment*; *undershirt*; *unicycle*; *uniform*; *urena lobata*; *ursinia*; *usb drive*; *usb interface*; *uvularia perfoliata*.

V (31 classes)

vachellia farnesiana; *vacuum cleaner*; *van*; *vancouveria hexandra*; *vase*; *vegetable oil*; *vehicle registration plate*; *velella velella*; *veneroida*; *verbascum virgatum*; *veronica wormskjoldii*; *vessel*; *vest*; *VGA interface*; *viguiera dentata*; *vincetoxicum nigrum*; *vinegar*; *viola adunca*; *viola blanda*; *viola macloskeyi*; *viola odorata*; *viola pedunculata*; *viola reichenbachiana*; *violet wood sorrel*; *violin*; *virginia oyster*; *virginia spring beauty*; *vitasoy*; *vitex agnus-castus*; *volleyball*; *volleyball net*.

W (85 classes)

wachendorfia paniculata; *waist chain*; *Waistcoat*; *walker*; *walking machine*; *wall socket*; *wall tent*; *wallet*; *wardrobe*; *wash lamp*; *watch*; *watch strap*; *water avens*; *water carrot*; *water castle*; *water chinquapin*; *water dropwort*; *water faucet*; *water gun*; *water hemlock*; *water lobelia*; *water shoe*; *water speedwell*; *water violet*; *water lily*; *watermelon*; *weapon*; *weathervane*; *webbing*; *weisswurst*; *well*; *welsh poppy*; *wheat straw painting*; *wheel*; *wheel chandelier*; *whelk*; *whirligig beetle*; *whisky*; *whistle*; *white bryony*; *white campion*; *white chocolate*; *white turnip*; *white wine*; *whorled loosestrife*; *wide-leg pants*; *wiener melange*; *wild geranium*; *wild parsley*; *wild pink*; *willow plait weaving*; *wimple*; *winchester defender shotgun*; *wind instrument*; *windbell*; *windbeutel*; *window*; *window covering*; *window frame*; *window valance*; *wine*; *wine cabinet*; *wine rack*; *wing nut*; *winter cress*;

wireless Bound Notebook; *wireless walkie-talkie*; *wok*; *wood anemone*; *wood carving*; *wood mint*; *wooden ball*; *wooden ladder*; *wooden toy train*; *workwear*; *worm*; *wrasse*; *wrench*; *wrist sweatband*; *wristband*; *writing brush*; *writing implement*; *wyethia angustifolia*; *wyethia mollis*; *würstelstand*.

X (2 classes)

xeranthemum; *xylorhiza orcuttii*.

Y (7 classes)

yard marker; *yellow chamomile*; *yellow foxglove*; *yogurt*; *yolk*; *youngia japonica*; *yurt*.

Z (3 classes)

zinnia elegans; *zipper*; *zwetschge*.

B VastFSS Construction Pipeline and Additional Annotation Examples

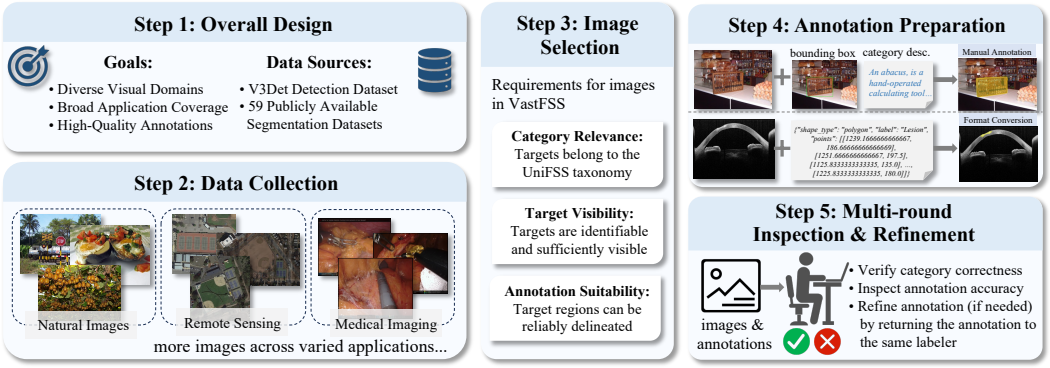


Figure 6: Construction pipeline of VastFSS, comprising overall design, data collection, image selection, annotation preparation, and multi-round inspection and refinement. Annotation preparation includes manual mask annotation and conversion of existing segmentation annotations.

As illustrated in Fig. 6, VastFSS is constructed through five key steps. **(i)** ***Overall design.*** We establish the construction goals of VastFSS, emphasizing a large category vocabulary, broad application coverage, and high-quality annotations. We also define the category taxonomy and identify suitable data sources. **(ii)** ***Data collection.*** We collect images and available annotations from the V3Det detection dataset and 59 additional publicly available datasets. **(iii)** ***Image selection.*** Reviewers with relevant research experience manually inspect the collected images for category relevance, target visibility, and annotation suitability. Only images meeting these criteria are retained. **(iv)** ***Annotation preparation.*** For images requiring pixel-level annotations, annotators manually delineate target regions using available bounding boxes and category descriptions as references. For images with existing segmentation annotations, we convert the corresponding annotations into the unified VastFSS format. **(v)** ***Multi-round inspection and refinement.*** Reviewers assess the annotations for semantic correctness and boundary accuracy. Annotations that do not meet our quality requirements are corrected or re-annotated and reviewed again. Additional annotation examples spanning a broad range of target categories and application contexts are presented in Fig. 7.

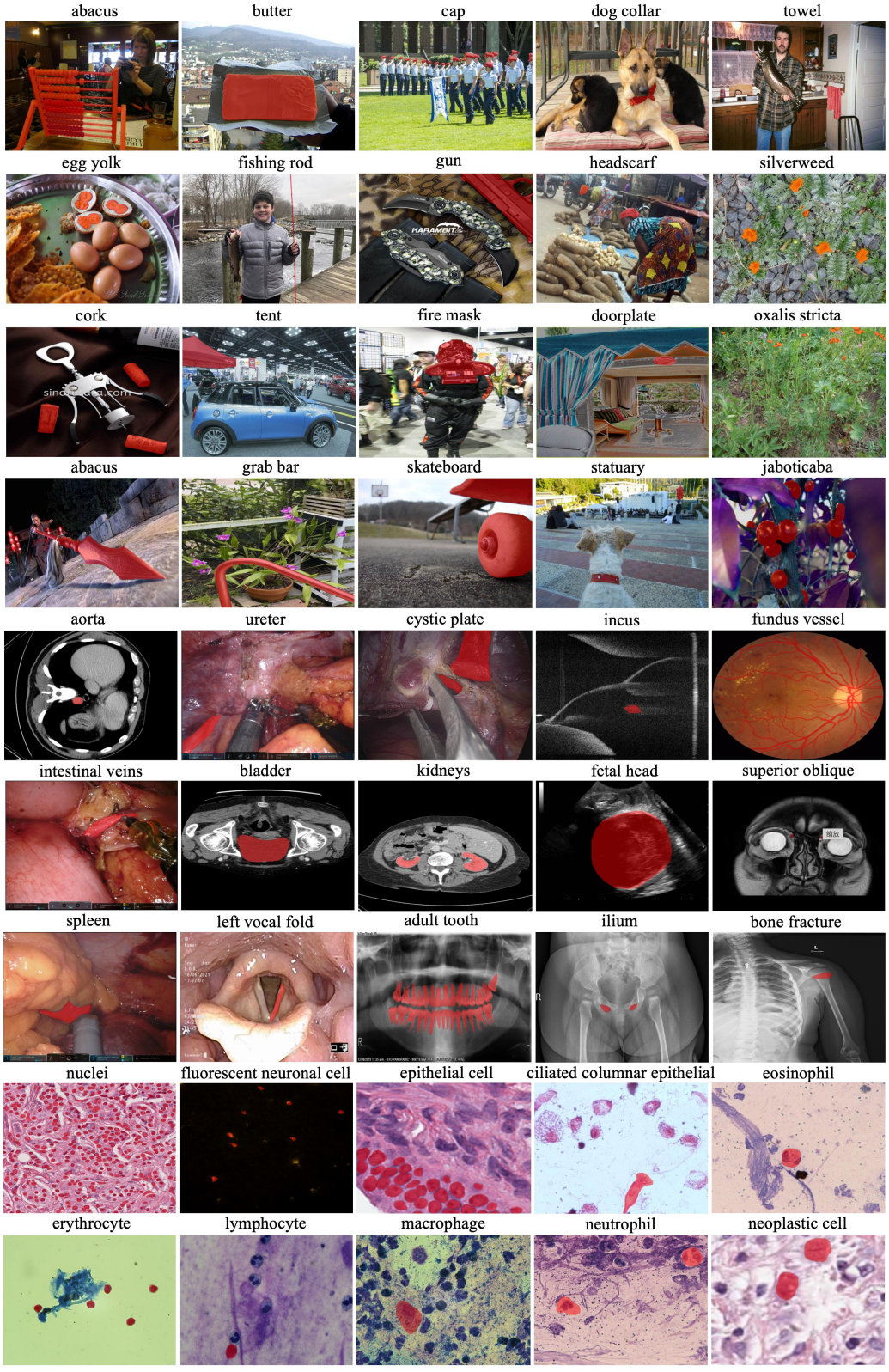


Figure 7: Additional annotation examples from VastFSS. Ground-truth masks are overlaid in red.

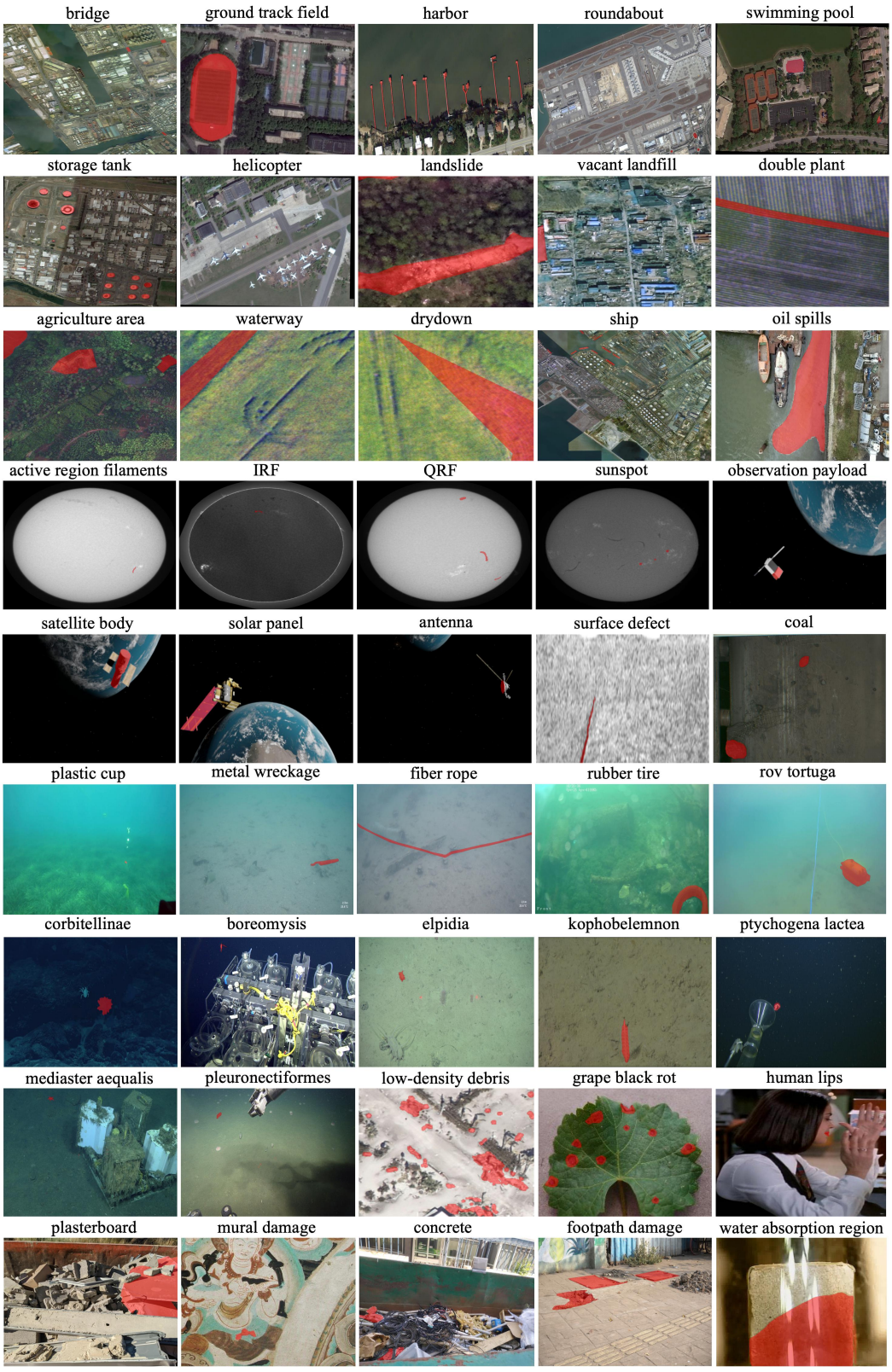


Figure 7: Additional annotation examples from VastFSS (continued).

C Detailed Architecture of the Prior Generation Module

As illustrated in Fig. 8, the prior generation module produces a support-guided spatial prior mask M_{prior} by measuring visual similarities between query features and support prototypes. Following (Xu et al., 2025), we use a frozen DINOv2 encoder (Oquab et al., 2024) to extract support and query features, denoted by $\widetilde{F}_s$ and $\widetilde{F}_q$, respectively.

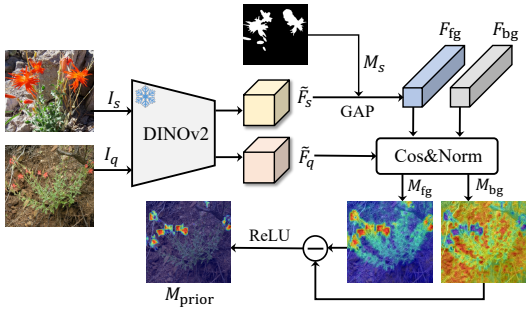


Figure 8: Architecture of the Prior Generation module.

The support mask M_s is first resized to match the spatial resolution of $\widetilde{F}_s$, yielding $\widetilde{M}_s$. Foreground and background prototypes are then constructed by applying global average pooling to the masked support features,

$$F_{\text{fg}} = \texttt{GAP}(\widetilde{F}_s \odot \widetilde{M}_s), \qquad F_{\text{bg}} = \texttt{GAP}(\widetilde{F}_s \odot (1 - \widetilde{M}_s)), \tag{11}$$

where $\texttt{GAP}(\cdot)$ denotes global average pooling and $\odot$ denotes element-wise multiplication, with the mask broadcast across feature channels.

We compute the cosine similarity between each prototype and the query feature at every spatial location, followed by min–max normalization,

$$M_{\text{fg}} = \texttt{Norm}\left(\texttt{CosSim}(\widetilde{F}_q, F_{\text{fg}})\right), \quad M_{\text{bg}} = \texttt{Norm}\left(\texttt{CosSim}(\widetilde{F}_q, F_{\text{bg}})\right), \tag{12}$$

where $\texttt{Norm}(\cdot)$ independently rescales each similarity map to $[0, 1]$. The spatial prior is obtained by subtracting the background similarity map from the foreground similarity map and applying ReLU to remove negative responses,

$$M_{\text{prior}} = \texttt{ReLU}(M_{\text{fg}} - M_{\text{bg}}). \tag{13}$$

This operation emphasizes query regions with stronger normalized foreground responses than background responses, providing spatial guidance for subsequent segmentation.

In the K-shot setting, we generate a prior map for each support sample and average the resulting maps to obtain the final spatial prior,

$$M_{\text{prior}} = \frac{1}{K} \sum_{k=1}^{K} M_{\text{prior}}^{(k)}, \tag{14}$$

where $M_{\text{prior}}^{(k)}$ denotes the prior generated from the k-th support sample and the query image.

D Details of Loss Function for Optimization

Our model is trained end-to-end using a text generation loss $\mathcal{L}_{\text{txt}}$ and a segmentation loss $\mathcal{L}_{\text{seg}}$. The overall training objective is defined as

$$\mathcal{L} = \lambda_{\text{txt}} \mathcal{L}_{\text{txt}} + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}}, \tag{15}$$

where λ_{txt} and λ_{seg} balance the two training objectives.

The text generation loss $\mathcal{L}_{\text{txt}}$ is implemented using the standard autoregressive cross-entropy loss. Given the ground-truth text $\mathbf{y}_{\text{txt}}$ and its prediction $\widehat{\mathbf{y}}_{\text{txt}}$, it is defined as

$$\mathcal{L}_{\text{txt}} = \text{CE}\left(\widehat{\mathbf{y}}_{\text{txt}}, \mathbf{y}_{\text{txt}}\right). \tag{16}$$

For segmentation, the mask decoder predicts N_{pre} candidate masks and their corresponding confidence logits, denoted by $\{(\widetilde{M}_q^i, s_i)\}_{i=1}^{N_{\text{pre}}}$. During training, bipartite matching is performed between the candidate masks and the ground-truth query mask M_q. Let $i^\dagger$ denote the index of the candidate mask matched to M_q, and let $\{s_i^*\}_{i=1}^{N_{\text{pre}}}$ denote the corresponding target confidence scores derived from the matching results.

The mask-quality score loss is defined as

$$\mathcal{L}_{\text{score}} = \frac{1}{N_{\text{pre}}} \sum_{i=1}^{N_{\text{pre}}} \text{BCE}\left(\sigma(s_i), s_i^*\right), \tag{17}$$

where $\sigma(\cdot)$ denotes the sigmoid function. This loss encourages the predicted confidence scores to reflect the quality of the candidate masks.

The segmentation loss is then computed using the matched candidate mask and the mask-quality score loss according to

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{bce}} \,\text{BCE}\left(\widetilde{M}_q^{i^\dagger}, M_q\right) + \lambda_{\text{dice}} \,\text{Dice}\left(\widetilde{M}_q^{i^\dagger}, M_q\right) + \lambda_{\text{score}} \mathcal{L}_{\text{score}}, \tag{18}$$

where λ_{bce}, λ_{dice}, and λ_{score} denote the corresponding loss weights. During training, we set $\lambda_{\text{bce}} = 2.0$, $\lambda_{\text{dice}} = 0.5$, and $\lambda_{\text{score}} = 1.0$.

During inference, bipartite matching is not required. The candidate with the highest predicted confidence score is selected as the final query mask according to

$$i^* = \underset{1 \le i \le N_{\text{pre}}}{\arg\max}\; s_i, \qquad \widehat{M}_q = \widetilde{M}_q^{i^*}. \tag{19}$$

E Qualitative Analysis

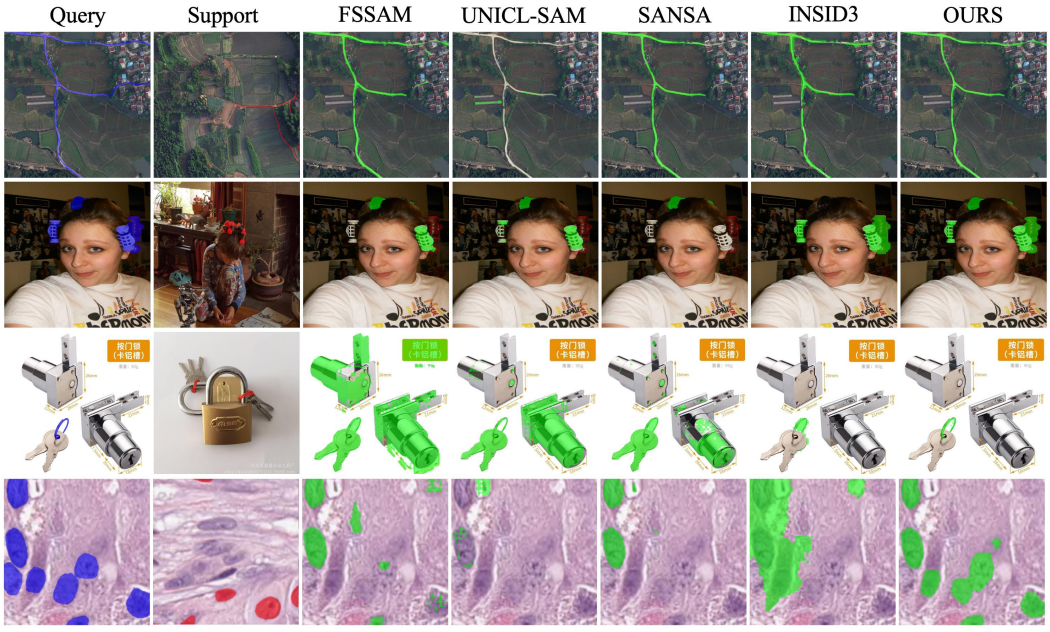


Figure 9: Visual comparison on Vast-FSS dataset under 1-shot setting. The ground-truth objects in query image are highlighted in blue, the referred objects in support image are highlighted in red, and the predicted objects in query image are highlighted in green.

To qualitatively evaluate MA-FSS, we compare its segmentation results on our VastFSS with those of representative methods (Xu et al., 2025; Sheng et al., 2025; Cuttano et al., 2025; 2026) in Fig. 9. In the illustrated examples, competing methods produce incomplete masks or include irrelevant background regions, particularly in the presence of occlusions and visually similar distractors. In contrast, MA-FSS produces more complete target masks with clearer boundaries and less background leakage. These qualitative results complement our quantitative findings, suggesting that the combination of target-aware representations and support-guided spatial priors helps distinguish the target from surrounding distractors in challenging segmentation scenarios.

Table 9: Quantitative comparisons on LVIS-92^i under the 1-shot setting. Columns 92^0–92^9 report the mIoU for each fold, and "Mean" denotes the average across all 10 folds. An asterisk ($*$) indicates training on LVIS-92^i. MA-FSS is trained on VastFSS and evaluated without additional training or fine-tuning on LVIS-92^i. The best available result in each column is highlighted in **bold**.

Method	1-shot										
	92^0	92^1	92^2	92^3	92^4	92^5	92^6	92^7	92^8	92^9	Mean
Matcher (Liu et al., 2024b) [ICLR'24]	31.4	30.9	33.7	38.1	30.5	32.5	35.9	34.2	33.0	29.7	33.0
SANSA* (Cuttano et al., 2025) [NeurIPS'25]	**48.4**	**48.3**	**51.5**	50.7	44.8	**50.1**	**51.1**	**50.5**	45.9	46.3	**48.8**
INSID3 (Cuttano et al., 2026) [CVPR'26]	38.6	39.6	41.5	43.9	37.1	42.5	42.8	41.3	41.8	37.3	40.6
MA-FSS (zero-shot transfer)	43.4	46.9	50.7	**57.0**	**48.1**	48.2	48.3	47.4	**48.5**	**46.4**	48.5

F Additional Results on LVIS-92^i

To further assess the transferability of MA-FSS, we directly evaluate the VastFSS-trained model on LVIS-92^i under the 1-shot setting, without additional training or fine-tuning on the target benchmark. As shown in Tab. 9, MA-FSS achieves a mean mIoU of 48.5%, outperforming Matcher and INSID3 by 15.5% and 7.9%, respectively. Notably, it performs competitively with SANSA, which is trained on LVIS-92^i, with a gap of only 0.3%, and achieves the best results on four of the ten folds. These results demonstrate the effectiveness of MA-FSS and suggest that VastFSS provides a useful training resource for few-shot segmentation beyond its own evaluation setting.

G Complementarity of Target-Aware Tokens and Spatial Prior

To investigate the complementary contributions of MLLM-derived target-aware tokens and support-guided spatial prior, we evaluate variants of MA-FSS that retain either component alone or combine both on VastFSS. As shown in Tab. 10, using only the spatial prior achieves a mean mIoU of 44.4% (❶), whereas using only the target-aware tokens achieves 54.0% (❷). Combining both increases the mean mIoU to 55.3% (❸), yielding gains of 10.9% and 1.3% over prior-only and token-only variants. The combined model outperforms both variants across all four folds, supporting the complementarity of the two sources of guidance. While the MLLM-derived tokens encode target semantics inferred from annotated support samples, the spatial prior provides explicit localization cues in the query image. These results suggest that spatial guidance complements the learned target representations, enabling more accurate query segmentation.

Table 10: Ablation on the support-guided spatial prior and MLLM-derived target-aware tokens.

	Prior	Tokens	Fold-0	Fold-1	Fold-2	Fold-3	Mean
❶	✓	-	44.2	47.2	43.5	42.7	44.4
❷	-	✓	52.8	55.5	53.0	54.7	54.0
❸	✓	✓	**54.0**	**56.9**	**54.4**	**55.8**	**55.3**

H Limitations and Future Work of MA-FSS

Despite its promising performance, MA-FSS remains challenged by scenarios with ambiguous object boundaries or targets that require implicit semantic reasoning. These difficulties may partly stem from the limited visual understanding and reasoning capabilities of the 2B-parameter MLLM used in our framework. However, segmentation quality also depends on the accuracy of the support-guided spatial prior and the ability of the mask predictor to resolve fine-grained boundaries. Future work will explore more capable MLLMs and improved spatial guidance to better handle these challenging cases, while balancing segmentation accuracy and computational cost.

I Ethical Considerations and Dataset Documentation

Ethical Considerations. We document and acknowledge all data sources used to construct VastFSS and adhere to their respective licenses and terms of use throughout data collection, processing, and integration. We will review any reported changes to source licenses or concerns regarding data use and take appropriate action, including updating access information or removing affected data where necessary.

Maintenance. We will maintain a GitHub repository for VastFSS to provide dataset documentation, access instructions, and updates. The repository will also serve as a channel for community feedback, including reports of annotation errors and data access issues. The authors will review this feedback and make necessary corrections to support the long-term availability and reliability of the benchmark.

Responsible Use. VastFSS is intended for research on generalizable few-shot segmentation. Its use is restricted to research purposes and remains subject to the licenses and terms of the source datasets. Users should be aware that data collection and source selection may introduce geographic, demographic, and domain imbalances. Benchmark performance should therefore not be assumed to reflect performance across all populations or application settings.

Potential Limitations. VastFSS is constructed from publicly available datasets, and its category vocabulary is therefore constrained by data availability. Targets associated with proprietary or access-restricted data may be underrepresented or absent, particularly in specialized applications. Despite its large vocabulary, VastFSS cannot fully represent the vast range of visual concepts encountered in real-world segmentation tasks. Future work could extend its category coverage through collaborations with relevant experts and institutions, subject to appropriate data access and usage permissions.