Adapting the Segment Anything Model Family for Specialized Domains: A Focused Review of Remote Sensing, Biological Imaging, and Video Segmentation

Vedant Misra
Master's Student, Computer Science and Engineering
The Pennsylvania State University
Advisor: Prof. Huijuan Xu

April 2026

Abstract

The Segment Anything Model (SAM) established promptable segmentation as a general-purpose foundation-model interface, but the literature from 2023 to 2026 demonstrates that this interface degrades in specialized domains where prompts are expensive, dense, temporally ambiguous, or operationally infeasible. This review argues that the central challenge in downstream SAM adaptation is not merely the distribution shift between natural and domain imagery—it is the structural mismatch between SAM's prompt-centric design and the operational realities of expert deployment. To substantiate this claim, we first synthesize the architectural evolution of the SAM family from SAM 1 through SAM 3.1, covering the introduction of streaming memory, Promptable Concept Segmentation, and Object Multiplexing. We then conduct a focused, domain-comparative analysis across three settings in which this mismatch is especially visible: remote sensing and change detection, microscopy and histopathology, and video analysis. In remote sensing, domain shift and bi-temporal semantics necessitate either automatic prompt generation with spectral priors or training-free bitemporal latent matching. In biological imaging, the impracticality of per-object prompting at cellular scale drives methods toward detector-driven auto-prompting, prompt-free decoding, and parameter-efficient fine-tuning with minimal trainable footprints. In video, temporal memory management, robustness under corruption, and multi-object scaling reshape the inference design space. Across all three domains, the strongest adaptations succeed not simply by fine-tuning the backbone but by redesigning the interaction regime—injecting domain structure, replacing manual prompts with learned alternatives, or restructuring inference for scale. The review synthesizes lessons from over forty representative methods and identifies three persistent open problems: prompt cost as an unmeasured evaluation variable, under-specified robustness and cross-dataset transfer protocols, and the false separation of computational efficiency from scientific quality.

1. Introduction

The Segment Anything Model (SAM) changed the practical language of image segmentation by formalizing mask prediction as a promptable foundation-model task [1]. Trained on the billion-mask SA-1B dataset, SAM demonstrated that a modular image encoder, prompt encoder, and lightweight mask decoder could support zero-shot interactive segmentation across highly diverse natural imagery. In natural-image settings, this interface is powerful because a small number of point or box prompts can often disambiguate an object of interest quickly and accurately. However, the literature that followed SAM's release makes clear that this interaction model does not transfer cleanly into specialized settings.

In remote sensing, the object of interest may be defined by semantic change across time rather than by a single static object boundary, and fine-grained geospatial structures do not align with the objectness prior learned from natural photographs [9]. In microscopy and histopathology, hundreds or thousands of tightly packed instances may appear in a single field of view, rendering per-object prompting impractical. In video, prompts must remain useful over long temporal horizons under corruption, motion, occlusion, and multi-object scale. These are not minor edge cases; they expose a structural mismatch between promptable segmentation as originally formulated and the operational realities of specialized deployment.

This review takes a focused rather than encyclopedic approach. Instead of cataloguing every downstream SAM application, it concentrates on three domains that most clearly illustrate the literature's central adaptation problem: remote sensing and change detection, biological imaging, and video analysis. Together, these areas reveal a common pattern: the best-performing methods succeed not simply because they adapt a strong pretrained backbone, but because they replace unrealistic prompting assumptions with domain-specific structure—through learned priors, automated prompt generation, prompt-free decoding, memory-aware scheduling, or shared-memory batching.

A substantial body of related work also exists in medical imaging, document understanding, scene text recognition, and underwater vision [21, 38, 39]. Those literatures are important, but they are used here only as supporting context for the broader argument that SAM adaptation has become an interaction-design problem as much as a representation-learning problem.

This review makes three explicit contributions. First, it reorganizes the recent literature around an evaluative question that is more informative than a simple model taxonomy: how does each method handle prompt burden under real deployment constraints? Second, it compares competing adaptation philosophies within each focal domain rather than offering a paper-by-paper inventory. Third, it argues that future evaluation protocols for specialized-domain SAM adaptation must measure prompt cost, robustness, and inference scaling explicitly rather than treating them as secondary engineering details.

2. Architecture of the SAM Family

Understanding the architectural progression of the SAM family is essential for contextualizing the downstream adaptations discussed in subsequent sections. The family is defined by a continuous expansion of its operational manifold—from static spatial priors to dynamic, multi-modal, and highly parallelized spatio-temporal reasoning.

2.1. SAM 1: Promptable Segmentation as a Foundation Task

SAM 1 introduced a tripartite modular architecture: a heavyweight image encoder, a flexible prompt encoder, and a lightweight mask decoder [1]. The image encoder, based on a Masked Autoencoder (MAE) pre-trained Vision Transformer (ViT), processes high-resolution imagery into dense spatial embeddings. The prompt encoder maps sparse prompts (points, boxes, text) and dense prompts (masks) into a common embedding space. The mask decoder, utilizing a two-way transformer block, facilitates attention between the image embedding and the prompt tokens. Trained on SA-1B—11 million licensed images paired with 1.1 billion automatically generated masks—SAM 1 achieved unprecedented zero-shot spatial generalization on natural imagery.

2.2. SAM 2: Streaming Memory and Temporal Propagation

SAM 2 extended the promptable interface to video object segmentation through a streaming-memory transformer [2]. A memory bank stored both spatial feature maps and high-level object representations, which were iteratively updated via self-attention to propagate segmentations through time. Training on SA-V—tens of millions of temporally consistent masklets across tens of thousands of video sequences—positioned temporal propagation and interactive spatial correction as the model's core competencies. For downstream adaptation, the key implication of SAM 2 is that a single prompt can now be amortized over many frames; but this also means that noisy or misplaced prompts can contaminate the memory and degrade long-horizon performance.

2.3. SAM 3: Promptable Concept Segmentation

SAM 3 reframed the objective from purely spatial demarcation to semantic comprehension via Promptable Concept Segmentation (PCS) [3]. Rather than requiring explicit coordinates, SAM 3 accepts short noun phrases (e.g., "yellow school bus") and visual exemplars, including hard negatives, to segment all instances of a concept globally. The architecture featured a decoupled image-level detector and a memory-based video tracker sharing a unified ViT backbone, augmented by a *presence head* that separated the recognition of a concept's existence from its precise spatial localization, dramatically reducing false-positive rates. Training on SA-Co introduced 4 million unique concept labels to encourage fine-grained semantic discrimination.

Table 1: Evolutionary capabilities of the SAM family.

Model	Core Architecture	Primary Dataset	Key Capability
SAM 1 (2023)	ViT encoder + prompt/mask decoder	SA-1B (1.1B masks)	Zero-shot 2D promptable segmentation.
SAM 2 (2024)	Streaming-memory transformer	SA-V (10M+ masklets)	Temporal propagation; interactive video correction.
SAM 3 (2025)	Decoupled detector/tracker + presence head	SA-Co (4M concepts)	Concept prompts (noun phrases, exemplars).
SAM 3.1 (2026)	Object Multiplexing (shared-memory batching)	SA-Co (refined)	16 objects per joint pass; 7× speedup at 128 objects.

2.4. SAM 3.1: Object Multiplexing and Inference Scalability

Released in March 2026, SAM 3.1 addressed the linear scaling bottleneck of SAM 3's multi-object video pipeline [4]. Under SAM 3, tracking N objects required N independent forward passes, causing massive computational redundancy. SAM 3.1 introduced *Object Multiplexing*: a hardware-aware shared-memory batching strategy in which a Multiplexer aggregates spatial and memory tracking data for up to 16 distinct objects from frame $T-1$ into fixed-capacity buckets, processes them in a single joint forward pass, and then a Demultiplexer separates the output into distinct identity masklets for frame T. This achieved a roughly 7× inference speedup when tracking 128 objects on a single NVIDIA H100 GPU and lifted throughput from 16 to 32 frames per second for medium object counts, while also yielding accuracy gains of +2.1 cgF1 on YT-Temporal-1B and +2.0 on MOSEv2 due to the joint inter-object context modeled during the shared pass.

3. Review Methodology

3.1. Selection Criteria

We prioritized peer-reviewed publications and pre-print technical reports from 2023 to 2026, aggregated from CVPR, NeurIPS, ICCV, MICCAI, and leading journals including *Nature Methods*, *Nature Communications*, and *IEEE JSTARS*. Papers were included if they either (i) introduced or formally documented a SAM-family release, or (ii) proposed a technically described downstream adaptation evaluated on domain-specific datasets with explicitly stated performance metrics. Works that used SAM merely as a static offline preprocessing step without methodological innovation in the segmentation pipeline were excluded. Each included paper was coded by application domain, adaptation mechanism, supervision requirement, prompt regime, and computational cost (trainable parameter fraction, auxiliary modules, and memory management strategies).

Table 2: Comparative framework applied across the three focal domains.

Domain	Primary SAM mismatch	Representative adaptation logic	What constitutes progress
Remote sensing & change detection	Multi-scale structure, small targets, semantic change across time	PEFT with domain priors; training-free bitemporal matching; Siamese CD pipelines	Reducing prompt reliance while improving cross-dataset transfer and suppressing pseudo-change
Microscopy & histopathology	Dense overlapping instances; infeasible per-object prompting; weak boundaries	Domain-specific fine-tuning; detector auto-prompting; prompt-free LoRA decoding	Maintaining instance quality without prompting hundreds of objects per image
Video analysis	Temporal drift, corruption sensitivity, multi-object scaling	Streaming-memory prompt schedules; concept prompting; shared-memory multiplexing	Balancing segmentation quality with prompt frequency, robustness, and throughput

3.2. Analytical Framework

This review evaluates each method through four lenses, applied consistently across all three focal domains.

1. **Domain mismatch:** What aspect of the domain breaks the assumptions of natural-image promptable segmentation?
2. **Prompt burden:** How much human prompting is required at inference, and is that level of interaction operationally realistic?
3. **Adaptation efficiency:** Does the method rely on full retraining, PEFT, detector-generated prompts, prompt-free decoding, or a deployment-time systems change?
4. **Evaluation realism:** Do the reported experiments measure transfer, robustness, prompt cost, and scale in a way that reflects actual use conditions?

4. Remote Sensing and Change Detection

4.1. Domain Challenges and Baseline Evaluation

Remote sensing is a revealing testbed because it compounds several distinct problems. Natural-image ViTs are trained on ground-level RGB photographs; overhead imagery differs in viewpoint, scale range, spectral statistics, and object definition. Ren et al. conducted one of the first systematic evaluations of SAM on overhead imagery and showed a mixed outcome: competitive transfer on

Table 3: RSAM-Seg performance on the 38-Cloud dataset [10].

Metric	**Score**
mIoU	0.7646
F1 Score	0.8152
Precision	0.8301
Recall	0.8396
Jaccard Index	0.7310
Overall Accuracy	0.9197

some tasks, but substantial degradation on buildings, solar arrays, and especially roads [9]. The deeper finding was conceptual: the mismatch is not only distributional but also structural. SAM's objectness prior is per-instance and prompt-driven; roads and land-cover regions are class-level phenomena that resist instance-centric prompting. Change detection adds a further complication—the target is a semantic difference across two temporal observations, making single-image prompting poorly matched to the task.

4.2. PEFT and Automatic Prompt Generation

One response to this mismatch is to preserve the SAM backbone while injecting remote sensing priors through lightweight modules. RSAM-Seg is a representative example [10]. Rather than relying on default prompting, it introduces two heavily engineered encoder-side modules: *Adapter-Scale*, injected into multi-head attention blocks using a residual MLP structure modulated by a scaling factor of 0.5 to preserve pre-trained attention maps; and *Adapter-Feature*, which supplements the network with high-frequency components (F_{hfc}) and geometric embeddings (F_{pe}) between ViT blocks.

Because geospatial targets such as road networks and building footprints rely on edge gradients rather than uniform color distributions, RSAM-Seg extracts F_{hfc} via the Fast Fourier Transform. The spatial tensor I is converted to the frequency domain via $f = \text{fft}(I)$, high-frequency coefficients are selectively retained with the low-frequency data shifted to the tensor center, and $I_{hfc} = \text{ifft}(f)$ reconstructs the enhanced spectral map. These priors drive an automatic prompt generator:

$$P_i = \text{MLP}_{up}(\text{GELU}(\text{MLP}_{tune}(F_{pe} + F_{hfc}))) \tag{1}$$

This eliminates manual prompting entirely at inference. Ablation studies on the 38-Cloud dataset confirm the necessity of the high-frequency prior: removing F_{hfc} drops mIoU by 0.0282 and F1 by 0.0267, and the full system achieves mIoU of 0.7646, F1 of 0.8152, and overall accuracy of 0.9197 (Table 3).

Related work extends this logic further. RS-SAM integrates multi-scale information for enhanced overhead segmentation [11]. RSPrompter frames remote-sensing instance segmentation as a prompt-generation problem layered on top of SAM, learning to produce semantically meaningful prompts

so that SAM can output category-aware instance masks [12]. GeoSAM combines automatically generated point prompts with text prompts for mobility infrastructure segmentation, reporting gains of at least 5% mIoU in both familiar and unseen geographic regions [13]. At a larger scale, RemoteSAM targets an Earth observation foundation model trained with a large image–text–mask data engine and a task-unified segmentation interface, indicating that the field is beginning to move beyond patching natural-image models toward remote-sensing-native pretraining [19].

4.3. Supervised Change Detection

A distinct line of work treats bi-temporal change detection as a dedicated adaptation problem. SAMCD adapts SAM features for high-resolution remote-sensing change detection using a convolutional adaptor and a task-agnostic semantic branch, demonstrating sample efficiency and strong supervised performance in bi-temporal settings [15]. PeftCD employs a weight-sharing Siamese framework with LoRA and adapter modules built on SAM 2 and other vision foundation model backbones, emphasizing cross-dataset generalization and the practical balance between accuracy and trainable parameter count [17]. SAM2-CD explicitly repurposes SAM 2's temporal architecture for the change detection task, showing that the streaming memory can be redirected toward bi-temporal feature comparison rather than temporal propagation [18]. More recent work such as FAEWNet combines distribution-aware spectral adaptation with edge-constrained warping for building change detection, arguing that when accurate boundaries and subtle structural change matter, explicit domain-aware feature adaptation remains superior to zero-shot transfer [20].

4.4. Training-Free Bitemporal Latent Matching

AnyChange offers the most conceptually distinctive response in the change detection literature by demonstrating that SAM's frozen latent space already contains change-sensitive structure [16]. Rather than retraining for change detection, AnyChange exploits SAM's object-centric capabilities by computing similarities over instance-level mask embeddings rather than raw pixel embeddings. Given a pair of registered bi-temporal images at times t and $t+1$, the frozen ViT encoder extracts dense embeddings z_t and z_{t+1}. For an automatically generated object proposal mask $m_{t,i}$, an instance-level mask embedding $x_{t,i}$ is isolated by averaging z_t over the non-zero spatial coordinates of the binary mask. Bidirectional similarity scores between $x_{t,i}$ and the corresponding region of z_{t+1} then identify semantic alterations without any temporal fine-tuning. AnyChange established a new state-of-the-art on the SECOND unsupervised change detection benchmark, surpassing previous deep learning architectures by up to 4.4% in F1 score, and supports sparse interactive refinement via a point query mechanism. This result is important because it demonstrates that training-free latent matching is a viable route for change detection, but it does not invalidate supervised adaptation: AnyChange is strong for zero-shot flexibility, while learned methods such as PeftCD and FAEWNet are stronger when pseudo-change suppression and precise boundary localization are required.

4.5. Synthesis: What Remote Sensing Teaches

Remote sensing exposes two limits of the original SAM formulation. The first is a *representation limit*: natural-image priors do not automatically resolve multi-scale geospatial structure or semantic change. The second is an *interaction limit*: when the target is a temporal difference, the idea of providing a simple spatial prompt to a single image is already poorly matched to the task. The strongest methods therefore either internalize prompt generation through domain-aware spectral and geometric priors, or shift the task toward bitemporal reasoning where prompting is sparse, automated, or replaced entirely. The largest open problem in this domain remains evaluation: many papers still report accuracy in ways that make prompt assumptions difficult to compare, and cross-dataset generalization under realistic temporal and sensor variation remains under-specified.

5. Microscopy, Biological Imaging, and Histopathology

5.1. Domain Challenges and the Prompt Burden at Scale

Microscopy and histopathology make the prompt burden problem unusually concrete. In natural-image segmentation, a user may want a handful of objects. In cell biology or pathology, a single field of view may contain hundreds or thousands of tightly packed instances with heavy overlap, weak boundaries, staining variation, and modality shifts across light microscopy, electron microscopy, and tissue imaging [29, 30]. A model that requires one prompt per instance is not merely inconvenient—it is structurally mismatched to the scale of the task.

5.2. Microscopy-Specific Fine-Tuning

A useful starting point is μSAM (*Segment Anything for Microscopy*), which extends SAM through microscopy-specific fine-tuning and delivers both interactive and automatic workflows for light and electron microscopy [29]. An important conclusion of that work is that a single universal microscope model did not emerge naturally from the original SAM architecture; instead, separate generalist models for light microscopy and electron microscopy were needed, implying that the microscopy problem is not solved by plug-and-play deployment but requires representation-level alignment to modality. μSAM additionally introduces volumetric segmentation, tracking support, and a napari plugin integrating annotation and retraining, making it a key anchor for any comparative analysis of biological-imaging adaptation. Recent benchmarking work confirms that μSAM features are especially effective for pixel and object classification when paired with lightweight classical learners, while SAM 2 features shine under attentive probing, suggesting that the optimal choice depends on the downstream task [37].

5.3. Automated Prompt Generation for Dense Instances

A second strategy is not to remove prompts but to move prompt generation from the human to an upstream detector. CellSAM is the clearest example: it combines SAM with CellFinder, a transformer-based detector built on the Anchor DETR framework, which formulates detection as a bipartite matching set prediction task [30]. Unlike dense pixel-wise frameworks (Mesmer, Cellpose) that rely on mutually exclusive pixel-to-cell assignment and fail on overlapping instances, Anchor DETR natively resolves dense clusters and partial occlusions without non-maximum suppression heuristics.

The integration uses a shared ViT-B backbone with an embedding dimension of 768. CellFinder decodes features to high-confidence bounding boxes, while a 2D convolutional model neck compresses embeddings from 768 to 256 dimensions for the SAM mask decoder. A two-stage fine-tuning protocol first trains the ViT backbone and CellFinder jointly on a detection objective; in the second stage, the ViT and SAM mask decoder are frozen and only the model neck and prompt encoder are updated using ground-truth boxes and segmentation labels. At inference, CellFinder generates boxes at a confidence threshold of 0.4, which are injected directly into the SAM prompt encoder, with mask logits evaluated at IoU 0.5. CellSAM reports lower segmentation error than Cellpose-generalist across all tested categories, human-level inter-rater agreement, and strong zero-shot performance on the LIVECell benchmark [30, 36]. Segment Any Cell occupies a similar space, reinforcing the pattern that automated prompting is a practical compromise—effective but not a final solution, because missed detections translate directly into missed segmentations [34].

PathoSAM extends this specialized-SAM philosophy to histopathology and reports state-of-the-art automatic and interactive nucleus instance segmentation across diverse nucleus benchmarks, though its semantic segmentation results still trail dedicated architectures such as CellViT [31].

5.4. Prompt-Free and Parameter-Efficient Adaptation

Recent work pushes further by removing the prompt interface entirely. Maqsood et al. (2026) introduce a prompt-free, lightweight SAM adaptation for H&E stained nuclei that discards the prompt encoder entirely and replaces it with multi-level encoder features coupled with a residual decoding module [35]. By fine-tuning only LoRA modules injected into the frozen SAM encoder, the method requires a minimal trainable footprint of 4.1 million parameters. Exhaustive evaluation across TNBC, MoNuSeg, and PanNuke demonstrates robust cross-dataset generalization, and the prompt-free design eliminates the compounded failure mode of auto-prompting pipelines: a missed detection by the bounding box generator guarantees a missed segmentation, whereas the prompt-free decoder has no upstream dependency.

UN-SAM similarly targets prompt-free nuclei segmentation through a self-prompt generation module and domain-adaptive encoder/decoder tuning, aiming at cross-domain generalization across tissue types and stain variation [32]. PTSAM demonstrates that strong adaptation need not require large-scale fine-tuning: by tuning only 2,048 parameters in the prompt tokens of the encoder and

decoder, it achieves competitive specialist performance in low-data regimes and remains viable with as few as 16 training images [33]. This result is especially significant for the argument of this review: it shows that for narrow, well-defined tasks, extreme PEFT can match domain-scale retraining.

For broader medical adaptation, H-SAM introduces a two-stage hierarchical decoding procedure using a LoRA-modified encoder [22]. Its Class-Balanced Mask-Guided Self-Attention (CMAttn) mechanism addresses highly unbalanced label distributions typical in clinical data, achieving a 4.78% improvement in average Dice similarity coefficient compared to existing prompt-free SAM variants for multi-organ segmentation using only 10% of available 2D training slices. S-SAM achieves comparable efficiency through Singular Value Decomposition (SVD)-based fine-tuning, updating only the largest singular values and lightweight LoRA-style residual matrices, restricting the trainable parameter footprint to a mere 0.4% of the original SAM architecture while enabling semantic label-name prompts [23].

5.5. Volumetric and Multi-Modal Extensions

Addressing the geometric mismatch of volumetric data, Tri-Plane Mamba introduces parameter-efficient adapters specifically engineered for depth-aware modeling in 3D CT scans [24]. Rather than processing volumes slice-by-slice (which induces severe inter-slice inconsistency), it projects features across orthogonal planes. MedSAM-2 reframes 3D segmentation entirely as continuous video tracking, using the SAM 2 streaming memory architecture alongside a self-sorting memory bank to propagate segmentations through a spatial volume from a single initial prompt [25]. At the other extreme, MedSAM demonstrates domain-scale retraining: fine-tuning the entire SAM architecture on 1,570,263 image-mask pairs across 10 imaging modalities and over 30 cancer types achieves a highly robust universal medical segmenter, but at the cost of catastrophic forgetting of general visual representations and massive compute requirements inaccessible to most clinical laboratories [21]. MV-SAM and AdaptSAM explore multi-view consistency and test-time adaptation respectively [27, 26].

5.6. Synthesis: What Biological Imaging Teaches

The biological imaging literature points to a three-part progression. First, microscopy-specific fine-tuning establishes that default SAM must be adapted at the representation level, and that modality matters—a single universal microscope model is insufficient. Second, auto-prompting methods such as CellSAM and PathoSAM scale dense-instance segmentation effectively, but transfer some failure risk upstream to the detector. Third, prompt-free methods such as UN-SAM, PTSAM, and the histopathology LoRA approach suggest that for the densest and most repetitive settings, eliminating prompts entirely may be the more coherent design choice. The field still lacks a standardized evaluation framework that compares interactive, auto-prompted, and prompt-free systems on equal footing; as a result, prompt cost is discussed rhetorically in most papers but rarely measured systematically.

6. Video Analysis, Scaling, and Evaluation

6.1. Streaming Memory and Prompt Scheduling

SAM 2 made video prompting a realistic foundation-model capability by combining spatial prompting with streaming memory [2]. In principle, temporal propagation should reduce interaction cost because a single prompt can be carried forward across frames. In practice, the specialized video literature shows that this promise is more fragile than it appears.

Shen et al. evaluated SAM 2 in surgical video environments subject to blood occlusion, rapid tool movement, specular highlights, and motion blur [41]. Their key finding is that continuous frame-wise prompting often *degrades* the streaming memory representation due to noise accumulation. Frame-sparse prompt scheduling—providing corrective spatial prompts only at optimized, infrequent intervals—frequently outperformed frame-wise prompting, because it allows the network to rely on its internal temporal momentum rather than forcing it to incorporate highly corrupted real-time spatial cues. This is a subtle but consequential result: more prompting is not always better prompting, and prompt schedules must be evaluated as part of the method, not assumed away. Pan et al. confirmed that zero-shot SAM 2 already performs competitively on standard video benchmarks such as DAVIS and YouTube-VOS, but that more demanding benchmarks—MOSE for cluttered scenes and LVOS for long-term sequences—reveal a larger gap between short-horizon and long-horizon robustness [42, 46, 47, 48, 49].

6.2. Long-Horizon Robustness

SAM2Long directly targets the long-horizon failure mode [43]. Rather than greedily trusting a single propagated mask, it treats memory selection as a training-free tree search: multiple segmentation pathways are maintained and video-level optimal paths are chosen retrospectively. This yields average gains of 3.0 points across head-to-head comparisons and up to 5.3 J&F on long-term benchmarks including SA-V and LVOS, illustrating that the dominant failure mode in video is often error accumulation in memory rather than a lack of spatial objectness.

SAMURAI addresses a related but distinct problem: zero-shot visual tracking under drift [44]. By adapting SAM 2 with motion-aware memory selection, it reports +7.1% AUC on $\text{LaSOT}_{\text{ext}}$ and +3.5% AO on GOT-10k without retraining the full model. These results suggest that in video the strongest directions combine challenging evaluation on long and complex sequences with explicit memory redesign. The BURST benchmark further broadens the picture by unifying recognition, segmentation, and tracking in video, exposing that temporal SAM adaptations optimized for one sub-task may not generalize to others [50].

6.3. Scaling-Efficient Video Adaptation

SAM-I2V makes a different argument: training cost should be part of the scientific comparison [45]. Rather than training a SAM 2-like model from scratch on large video corpora, it upgrades an

image-based SAM to promptable video segmentation through temporal feature integration, memory filtering, and a memory-as-prompt strategy, reporting over 90% of SAM 2 performance at only 0.2% of the training cost. This is an important methodological point: a method that is slightly weaker on a benchmark but reproducible with modest compute may be more scientifically and practically valuable than one that depends on inaccessible training pipelines.

6.4. SAM 3 Concept Prompting and SAM 3.1 Object Multiplexing

SAM 3 extends the interaction space by introducing concept prompts (noun phrases and image exemplars) that can unify segmentation, detection, and tracking more flexibly than purely spatial prompts [3]. For specialized domains, the significance of this is architectural: the family is already shifting toward richer interaction abstractions, suggesting that the long-term trajectory moves away from the original point/box-centric interface. SAM 3.1 then made multi-object inference scaling a first-order design concern through Object Multiplexing [4]. This is a systems-level change with scientific implications: multi-object video segmentation in realistic scenes cannot be meaningfully evaluated only by per-object mask quality if the computational pipeline scales quadratically with object count. Applications including traffic monitoring, surgical scenes, and biological cell tracking care jointly about throughput, object count, and latency, making Object Multiplexing directly relevant to downstream deployment.

6.5. Evaluation Benchmarks and Protocols

The choice of benchmark in video is inseparable from the scientific claim. DAVIS 2017 and YouTube-VOS remain canonical for short-to-medium sequence evaluation [46, 47]. MOSE adds crowded scenes and heavy occlusions, demonstrating that methods scoring near 90 J&F on DAVIS often fall substantially lower in cluttered settings [48]. LVOS introduces substantially longer sequences and repeated reappearance events, making it a stronger stress test for memory architecture [49]. A method that helps on DAVIS but not on LVOS is not solving long-horizon memory failure. This distinction matters because several published video SAM adaptations are evaluated exclusively on short benchmarks, making their practical contribution difficult to assess.

6.6. Synthesis: What Video Adaptation Teaches

The video literature shows that specialized-domain SAM adaptation must be evaluated on at least three axes simultaneously: segmentation quality, prompt schedule, and inference scale. Temporal memory makes prompting less expensive in principle, but introduces failure modes related to drift, corruption, and over-correction that interact with prompt frequency in non-obvious ways. Meanwhile, multi-object scaling turns inference design into part of the modeling problem. The critical gap is that these variables are almost always evaluated separately; a standardized protocol that jointly reports accuracy, prompt frequency, object count, and throughput over long horizons does not yet exist.

7. Cross-Domain Synthesis and Open Problems

Across remote sensing, biological imaging, and video, one pattern is unmistakable: the most successful adaptations preserve useful SAM-family priors while redesigning the interaction interface around the domain. In remote sensing, this means incorporating multi-scale and bitemporal structure so that change detection is not treated as ordinary single-image segmentation. In microscopy, it means admitting that per-object prompts are often infeasible and replacing them with detector-generated or prompt-free alternatives. In video, it means recognizing that prompt scheduling and multi-object scale are part of what determines whether a model remains usable in production. This pattern suggests that the main question for future research is not only how much of the backbone to fine-tune, but how to design a segmentation interface that respects the structure of specialized data and the realities of expert workflows.

7.1. Prompt Cost Must Become a Measured Quantity

The literature routinely describes prompt burden as a major limitation, but few papers evaluate it in a comparable, quantitative way. Some studies assume oracle bounding boxes; some use interactive point clicks; some automate prompting with detectors; some eliminate prompts entirely. These differences are often reported in adjacent papers as though the downstream tasks were directly comparable. Future benchmarks should report prompt frequency, prompt type, correction budget, and annotation-time proxies explicitly—for example, performance as a function of the number of user interactions (an interaction curve), so that the practical tradeoff between automation and accuracy can be measured rather than asserted.

7.2. Robustness and Transfer Remain Underspecified

A second open problem is robustness under realistic distribution shift. Remote-sensing papers often claim improved generalization but still evaluate on narrow dataset collections. Biological imaging papers frequently demonstrate strong in-domain performance while varying the prompt regime to such a degree that cross-paper comparison is unreliable. Video evaluations may measure segmentation under corruption *or* measure throughput, but rarely both together. The field needs protocols that test whether an adaptation remains stable when prompt regime, sensor characteristics, object density, or temporal conditions change—conditions that reflect real deployment scenarios.

7.3. Efficiency Is Not Separate from Scientific Quality

The review also demonstrates that efficiency and scaling should not be treated as secondary engineering concerns. PEFT in remote sensing and histopathology, auto-prompting in dense microscopy, and Object Multiplexing in video all change what kinds of deployment are possible. In specialized domains, this matters scientifically: a method that cannot scale to realistic object counts or prompt budgets does not solve the task in its intended setting. Future work should treat

throughput, memory, and trainable parameter footprint as part of the adaptation argument rather than as footnotes. SAM-I2V's demonstration that 0.2% of training cost recovers 90% of SAM 2 performance is one example of the kind of efficiency-versus-quality analysis that should become standard [45].

7.4. Toward Domain-Sensitive Foundation Interfaces

The broader implication is that "segment anything" is no longer a stable or sufficient downstream objective. Specialized-domain adaptation increasingly points toward a family of domain-sensitive segmentation interfaces: bitemporal segmentation for Earth observation, dense-instance or prompt-free segmentation for biology, and memory-aware or concept-aware segmentation for video. The challenge for the next generation of models is to support these interfaces without losing the flexibility that made foundation models attractive in the first place. Systems that output calibrated uncertainty estimates and actively request prompts only where confidence is low could substantially reduce annotation burden while maintaining the human-in-the-loop safety required in clinical and Earth-observation contexts.

8. Conclusion

The 2023–2026 literature on SAM adaptation shows that specialized-domain performance cannot be understood as a simple matter of transferring a stronger segmentation backbone into a new dataset. Across remote sensing change detection, microscopy and histopathology, and video analysis, the deeper problem is the mismatch between SAM's original prompt-centric interface and domains in which prompts are dense, temporal, ambiguous, or expensive to supply manually. The most successful adaptations therefore do more than fine-tune features: they redesign the interaction regime through domain priors, automated prompting, prompt-free decoding, temporal memory strategies, or scalable multi-object inference.

Concretely, remote sensing has moved from manual prompting toward spectral-prior PEFT and bitemporal latent matching; biological imaging has progressed through microscopy-specific fine-tuning, detector-driven auto-prompting, and increasingly prompt-free lightweight decoders; and video has seen temporal memory redesign and production-grade multi-object inference become the central engineering frontier. Architecturally, SAM 3.1's Object Multiplexing demonstrates that inference scalability is now a first-order modeling concern, and SAM-I2V's efficiency results show that training cost must enter the comparative evaluation.

The three open problems identified here—unmeasured prompt cost, underspecified robustness, and the false separation of efficiency from quality—point to a single overarching challenge. The future of specialized-domain segmentation with foundation models will depend less on preserving the original "segment anything" interface and more on building domain-sensitive interfaces that make prompting sparse, internalized, or unnecessary. That is the strongest common lesson shared by the remote-sensing, biological-imaging, and video literatures, and it is the most useful lens for evaluating

Table 4: Representative adaptations reviewed in this paper and the specific lesson each contributes to the central argument.

Method	**Adaptation type**	**Prompt requirement**	**Main lesson**
RSAM-Seg [10]	PEFT with remote-sensing priors; FFT high-frequency priors	No manual prompts	Domain-aware spectral priors enable fully automated remote-sensing segmentation.
AnyChange [16]	Training-free bitemporal latent matching	Sparse or query-based	Change detection can be induced from SAM representations without retraining.
PeftCD [17]	Siamese PEFT with LoRA on SAM 2 backbone	Prompt-free end-task	PEFT improves cross-dataset change detection robustness.
μSAM [29]	Microscopy-specific fine-tuning; new decoder	Interactive and automatic	A domain-generalist microscopy baseline is essential before specialized patches are compared.
CellSAM [30]	Detector-generated auto-prompting (Anchor DETR)	Automated box prompts	Auto-prompting scales dense-instance segmentation but transfers failure risk upstream.
H-SAM [22]	Two-stage hierarchical LoRA decoding; CMAttn	Prompt-free	Structured hierarchical priors achieve strong medical segmentation without expert prompts.
Prompt-free histopathology [35]	LoRA-based prompt-free decoding; 4.1M params	None	In the densest settings, prompt elimination is more coherent than prompt automation.
PTSAM [33]	Prompt tuning (2,048 parameters)	Prompt-free task specialization	Extreme PEFT can match domain-scale retraining in low-data regimes.
Surgical SAM 2 [41]	Prompt-schedule evaluation under corruption	Sparse corrective prompts	Prompt frequency is an evaluation variable; more prompts can hurt under noise.
SAM2Long [43]	Training-free memory tree over SAM 2	Standard video prompts	Long-video robustness requires memory redesign; short-benchmark scores are insufficient.
SAMURAI [44]	Motion-aware memory selection	Standard video prompts	Motion priors reduce drift without full retraining.
SAM-I2V [45]	Temporal upgrader + memory-as-prompt	Promptable video	90% of SAM 2 performance at 0.2% training cost; efficiency is part of the argument.
SAM 3.1 Object Multiplex [4]	Shared-memory multi-object batching	Concept or spatial prompts	Inference scaling is part of the model design, not a deployment afterthought.

the next wave of SAM-family adaptations.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.

[2] N. Ravi et al., "SAM 2: Segment Anything in Images and Videos," *arXiv preprint arXiv:2408.00714*, 2024.

[3] N. Carion et al., "SAM 3: Segment Anything with Concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[4] Meta / facebookresearch, "SAM 3.1 Release Notes: Object Multiplex," GitHub release notes, March 27, 2026. https://github.com/facebookresearch/sam3/blob/main/RELEASE_SAM3p1.md.

[5] Facebook Research, "SAM 2 repository and release notes," GitHub. Available: https://github.com/facebookresearch/sam2. Accessed Apr. 2026.

[6] L. Ke, M. Ye, M. Danelljan, Y. Liu, Y.-W. Tai, C.-K. Tang, and F. Yu, "Segment Anything in High Quality," *arXiv preprint arXiv:2306.01567*, 2023.

[7] T. Chen, X. Xiang, L. Liu, H. Wu, D. Yang, and W. Yang, "SAM-Adapter: Adapting Segment Anything in Underperformed Scenes," in *Proc. IEEE/CVF ICCV Workshops*, 2023.

[8] Y. Xiong et al., "EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[9] S. Ren, F. Luzi, S. Lahrichi, K. Kassaw, L. M. Collins, K. Bradbury, and J. M. Malof, "Segment Anything, From Space?," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[10] J. Zhang, X. Yang, R. Jiang, W. Shao, and L. Zhang, "RSAM-Seg: A SAM-based Approach with Prior Knowledge Integration for Remote Sensing Image Semantic Segmentation," *Remote Sensing*, vol. 17, no. 4, p. 590, 2025.

[11] E. Zhang et al., "RS-SAM: Integrating Multi-Scale Information for Enhanced Remote Sensing Image Segmentation," in *Proc. Asian Conference on Computer Vision (ACCV)*, 2024.

[12] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation based on Visual Foundation Model," *arXiv preprint arXiv:2306.16269*, 2023.

[13] R. I. Sultan, C. Li, H. Zhu, P. Khanduri, M. Brocanelli, and D. Zhu, "GeoSAM: Fine-tuning SAM with Multi-Modal Prompts for Mobility Infrastructure Segmentation," *arXiv preprint arXiv:2311.11319*, 2023.

[14] D. Wang et al., "SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model," in *NeurIPS Datasets and Benchmarks*, 2023.

[15] L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone, "Adapting Segment Anything Model for Change Detection in HR Remote Sensing Images," *arXiv preprint arXiv:2309.01429*, 2024.

[16] Z. Zheng, Y. Zhong, L. Zhang, and S. Ermon, "Segment Any Change (AnyChange)," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[17] S. Dong, Y. Hu, L. Wang, G. Chen, and X. Meng, "PeftCD: Leveraging Vision Foundation Models with Parameter-Efficient Fine-Tuning for Remote Sensing Change Detection," *arXiv preprint arXiv:2509.09572*, 2025.

[18] Y. Qin, C. Wang, Y. Fan, and C. Pan, "SAM2-CD: Remote Sensing Image Change Detection with SAM2," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[19] L. Yao et al., "RemoteSAM: Towards Segment Anything for Earth Observation," *arXiv preprint arXiv:2505.18022*, 2025.

[20] Y.-C. Li, S. Lei, Y.-T. Zhao, H.-C. Li, J. Li, and A. Plaza, "SAM-Based Building Change Detection with Distribution-Aware Fourier Adaptation and Edge-Constrained Warping," *arXiv preprint arXiv:2504.12619*, 2025.

[21] J. Ma et al., "Segment Anything in Medical Images (MedSAM)," *Nature Communications*, vol. 15, p. 654, 2024.

[22] Z. Cheng et al., "Unleashing the Potential of SAM for Medical Adaptation via Hierarchical Decoding (H-SAM)," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[23] J. N. Paranjape et al., "S-SAM: SVD-based Fine-Tuning of Segment Anything Model for Medical Image Segmentation," in *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

[24] H. Wang et al., "Tri-Plane Mamba: Efficiently Adapting Segment Anything Model for 3D Medical Images," in *Proc. MICCAI*, 2024.

[25] J. Zhu et al., "Medical SAM 2: Segment Medical Images as Video via Segment Anything Model 2," *arXiv preprint arXiv:2408.00874*, 2024.

[26] R. Schön et al., "Adapting the Segment Anything Model During Usage in Novel Situations," in *CVPR Workshops*, 2024.

[27] Anonymous, "MV-SAM: Multi-view Promptable Segmentation using Pointmaps for 3D Consistency," *arXiv preprint arXiv:2601.17866*, 2026.

[28] G. Xu et al., "Adapting Segment Anything Model 3 for Lesion Segmentation," *arXiv preprint arXiv:2603.25945*, 2026.

[29] A. Archit et al., "Segment Anything for Microscopy (μSAM)," *Nature Methods*, vol. 22, pp. 579–591, 2025.

[30] M. Marks et al., "CellSAM: a Foundation Model for Cell Segmentation," *Nature Methods*, vol. 22, no. 12, pp. 2585–2593, 2025.

[31] T. Griebel, A. Archit, and C. Pape, "Segment Anything for Histopathology (PathoSAM)," *arXiv preprint arXiv:2502.00408*, 2025.

[32] Z. Chen, Q. Xu, X. Liu, and Y. Yuan, "UN-SAM: Universal Prompt-Free Segmentation for Generalized Nuclei Images," *arXiv preprint arXiv:2402.16663*, 2024.

[33] T. Piater, B. Barz, and A. Freytag, "Prompt-Tuning SAM: From Generalist to Specialist with only 2,048 Parameters and 16 Training Images," *arXiv preprint arXiv:2504.16739*, 2025.

[34] S. Na et al., "Segment Any Cell: A SAM-based Auto-prompting Fine-tuning Framework for Nuclei Segmentation," *arXiv preprint arXiv:2401.13220*, 2024.

[35] M. H. Maqsood, Y. Zhu, A. Lam, G. Dagnaw, X. Yin, and A. W.-C. Liew, "Prompt-Free Lightweight SAM Adaptation for Histopathology Nuclei Segmentation with Strong Cross-Dataset Generalization," *arXiv preprint arXiv:2603.20326*, 2026.

[36] C. Edlund et al., "LIVECell—A Large-Scale Dataset for Label-Free Live Cell Segmentation," *Nature Methods*, vol. 18, pp. 1038–1045, 2021.

[37] A. Archit, F. Wagner, T. Griebel, and C. Pape, "Evaluating Vision Foundation Models for Pixel and Object Classification in Microscopy," *arXiv preprint arXiv:2603.19802*, 2026.

[38] X.-H. Li, F. Yin, and C.-L. Liu, "DocSAM: Unified Document Image Segmentation via Query Decomposition and Heterogeneous Mixed Learning," in *Proc. IEEE/CVF CVPR*, 2025.

[39] Y. Hong et al., "WaterSAM: Adapting SAM for Underwater Object Segmentation," *Journal of Marine Science and Engineering*, vol. 12, no. 9, p. 1616, 2024.

[40] S. Lian and H. Li, "Evaluation of Segment Anything Model 2: The Role of SAM2 in the Underwater Environment," *arXiv preprint arXiv:2408.02924*, 2024.

[41] Y. Shen et al., "Performance and Non-adversarial Robustness of the Segment Anything Model 2 in Surgical Video Segmentation," *arXiv preprint arXiv:2408.04098*, 2024.

[42] F. Pan, Y. Sun, Y. Xing, X. Zhang, H. Xie, and G. Li, "Video Object Segmentation via SAM 2," *arXiv preprint arXiv:2408.10125*, 2024.

[43] S. Ding et al., "SAM2Long: Enhancing SAM 2 for Long Video Segmentation with a Training-Free Memory Tree," in *Proc. IEEE/CVF ICCV*, 2025.

[44] C.-Y. Yang, H.-W. Huang, W. Chai, Z. Jiang, and J.-N. Hwang, "SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory," *arXiv preprint arXiv:2411.11922*, 2024.

[45] H. Mei, P. Zhang, and M. Z. Shou, "SAM-I2V: Upgrading SAM to Support Promptable Video Segmentation with Less than 0.2% Training Cost," in *Proc. IEEE/CVF CVPR*, 2025.

[46] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 DAVIS Challenge on Video Object Segmentation," *arXiv preprint arXiv:1704.00675*, 2017.

[47] N. Xu et al., "YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark," *arXiv preprint arXiv:1809.03327*, 2018.

[48] H. Ding, C. Liu, S. He, X. Jiang, P. H. S. Torr, and S. Bai, "MOSE: A New Dataset for Video Object Segmentation in Complex Scenes," *arXiv preprint arXiv:2302.01872*, 2023.

[49] L. Hong et al., "LVOS: A Benchmark for Large-Scale Long-Term Video Object Segmentation," *arXiv preprint arXiv:2404.19326*, 2024.

[50] A. Athar, J. Luiten, P. Voigtlaender, T. Khurana, A. Dave, B. Leibe, and D. Ramanan, "BURST: A Benchmark for Unifying Object Recognition, Segmentation and Tracking in Video," in *Proc. IEEE/CVF WACV*, 2023.