Adaptive Parameter-Efficient Fine-Tuning for Open-World Video Object Segmentation in Dynamic Scenes

Rowan J. Lindgren
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
lindgren847@unh.edu

Abstract

The emergence of vision foundation models has radically transformed video object segmentation, yet their deployment in open-world dynamic scenes presents systemic challenges that extend far beyond model accuracy. This work systematically examines the design of adaptive parameter-efficient fine-tuning frameworks tailored to open-world video object segmentation under non-stationary conditions, emphasizing architectural decisions, infrastructure orchestration, robustness governance, and socio-technical policy dimensions. We argue that the current generation of parameter-efficient adapters, low-rank adaptors, and prompt-based tuning modules, while reducing the computational footprint of fine-tuning, lacks the architectural reflexivity needed to continuously reconfigure learnable parameters in response to distributional shifts, novel object categories, and resource-constrained edge environments. We propose a conceptual architecture in which a meta-controller monitors scene dynamics, uncertainty signals, and hardware telemetry to decide online which subsets of parameters should be updated, and at what granularity, without resorting to full-model retraining. Through this lens, the paper analyzes structural trade-offs involving latency, memory, energy consumption, and fairness, drawing on a broad body of evidence from video segmentation benchmarks, green AI studies, and algorithmic fairness research. Further, we explore the deployment ecosystems required to sustain such adaptive systems across geographies with heterogeneous regulatory regimes, addressing privacy, data sovereignty, and accountability gaps. The discussion extends to governance mechanisms that can audit adaptation trajectories and prevent discriminatory feedback loops in safety-critical applications. By synthesizing cross-domain insights from computer vision, systems engineering, and policy studies, the paper offers a roadmap for building robust, sustainable, and ethically-grounded segmentation infrastructures that can gracefully handle the unbounded variability of open-world dynamic scenes.

Keywords

parameter-efficient fine-tuning, open-world video object segmentation, dynamic scenes, adaptive systems, system design, robustness, governance.

1. Introduction

The confluence of large-scale vision transformers and internet-scale pre-training has inaugurated a new era in video object segmentation, characterized by foundational models that demonstrate remarkable zero-shot transfer capabilities. The Segment Anything paradigm, instantiated in image models and subsequently extended to video streams through architectures like SAM 2, has shown that a monolithic neural network trained on millions of annotated masks can generalize to scenes with radically different visual statistics [1] [2]. Concurrently, specialized memory-based architectures such as XMem and the family of associating-objects transformers, including AOT and its decoupled counterpart DeAOT, have

pushed the boundaries of semi-supervised and unsupervised video object segmentation by hierarchically storing and reading spatiotemporal features across long sequences [3] [4] [5]. These advances have largely been evaluated under closed-world assumptions, where the categories of objects to be segmented are known in advance and the environmental conditions are relatively stable. However, when such models are deployed into the uncontrolled milieu of autonomous driving, disaster response robotics, or aerial surveillance, they must contend with open-world instantiations that introduce novel object categories, severe lighting variations, and continuous domain drift, all while operating under stringent resource budgets and minimal human oversight.

Traditional fine-tuning strategies that update all model parameters are unsuitable for these operational contexts because they incur prohibitive computational costs, risk catastrophic forgetting when sequentially adapted to new distributions, and fail to provide fine-grained control over which representations change. In response, a rapidly maturing body of work on parameter-efficient fine-tuning, encompassing low-rank adaptors (LoRA), lightweight adapter layers, and prefix- or prompt-tuning techniques, has demonstrated that competitive or even superior downstream performance can be achieved by introducing and updating only a tiny fraction of a model's weights [6] [7] [8]. When these methods are applied to video segmentation, they dramatically lower the barrier for per-instance or per-scene customization, opening the possibility of on-the-fly model personalization on edge devices. Yet, existing parameter-efficient fine-tuning pipelines are almost entirely static: the choice of which modules to train and the rank or capacity of the injected parameters are fixed at design time, leaving the system unable to renegotiate its adaptation budget when confronted with a sudden appearance of unfamiliar object categories or a rapid degradation in compute availability.

The open-world video object segmentation problem intensifies these demands precisely because the set of semantic concepts, motion patterns, and occlusions cannot be exhaustively enumerated during offline pre-training. Benchmarks such as Unidentified Video Objects (UVO), BURST, Occluded Video Instance Segmentation (OVIS), and Tracking Any Object (TAO) were explicitly constructed to evaluate segmentation and tracking systems on categories not seen during training, and they frequently feature crowded scenes with heavy inter-object occlusion and abrupt camera motion [10] [11] [12] [13]. These settings reveal fundamental brittleness even in state-of-the-art models that rely on fixed visual vocabularies. Researchers have thus sought to infuse vision-language alignment, most notably via CLIP embeddings, and open-vocabulary segmentation frameworks such as OpenSeg to broaden the semantic coverage of segmentation heads without exhaustive annotation [9] [15]. Meanwhile, universal segmentation architectures like Mask2Former and the YouTube-VIS suite have supplied unified pipelines that can handle instance, semantic, and panoptic outputs, though they were not originally designed for online adaptation in resource-constrained dynamic environments [14] [16]. These developments lay the groundwork for a new system class—an adaptive parameter-efficient fine-tuning framework that is structurally aware of its own computational and memory envelope and can selectively mobilize adaptation capacity where and when it is most needed.

The contribution of this paper is a system-level analytical framework for designing, deploying, and governing such adaptive architectures. Rather than proposing a single algorithm or numeric objective, we articulate a layered design space that spans the interaction between a scene-dynamics observer, a meta-policy for parameter selection, a resource scheduler, and a continuous risk assessment module. This design space is examined against a backdrop of

empirical evidence drawn from the parameter-efficient fine-tuning literature and from operational experience with large-scale video analytics systems. The paper is organized as follows. Section 2 maps the landscape of open-world video segmentation systems and identifies the core requirements that any adaptive fine-tuning framework must satisfy. Section 3 explicates the internal architecture of an adaptive parameter controller, comparing different granularities of adaptation and discussing how memory-aware fine-tuning strategies can be integrated. Section 4 turns to the infrastructure and deployment considerations that govern real-world viability, including edge-cloud partitioning, energy sustainability, and continuous learning pipelines. Section 5 analyzes the robustness and fairness dimensions of adaptive tuning in dynamic scenes, focusing on failure mode taxonomies and bias amplification risks. Section 6 broadens the analysis to governance, policy, and socio-technical implications, and Section 7 concludes with forward-looking recommendations.

2. Conceptual Foundations and System Requirements

A comprehensive understanding of the system requirements for adaptive fine-tuning demands that we first situate video object segmentation within the broader lineage of visual recognition architectures. Classical video object segmentation pipelines largely operated under the aegis of semi-supervised labeling, where the target object is defined by a single mask in the first frame and must be tracked throughout the video. This setup, while clean, is insufficient for open-world settings, where the very notion of a target object may emerge only midway through a sequence, or where multiple task definitions coexist in a single deployment. The shift toward open-world segmentation entails a move from discriminatively trained category-specific detectors to task-agnostic segmentation primitives that can be composed on the fly. Foundation models such as SAM have instantiated this capability for static images, and their temporal extensions to video have been realized by incorporating lightweight memory banks that store spatiotemporal tokens from past frames, as evidenced in the SAM 2 architecture [2]. Yet, the transition from static zero-shot segmentation to temporally coherent open-world segmentation reveals a crucial gap: the memory representations, attention patterns, and mask decoding parameters that were optimal for the pre-training data manifold often misalign with the long-tailed distributions encountered in the wild.

The core requirement for any system that intends to operate reliably in open-world dynamic scenes is the capacity for continual, context-sensitive reconfiguration of its internal representations without catastrophic interference with previously acquired capabilities. This requirement is articulated along multiple axes. From a computational standpoint, the reconfiguration must be economical, consuming only a small fraction of the floating-point operations and memory bandwidth demanded by full fine-tuning. From a temporal standpoint, the adaptation must be sufficiently fast to keep pace with a 30-frames-per-second stream while maintaining temporal consistency of masks across frames. From a safety standpoint, the reconfiguration must be auditable and reversible: in domains such as medical image assistance or autonomous navigation, it is unacceptable for a model to silently rewrite its segmentation policy based on a few anomalous frames. These three axes—efficiency, latency, and controllability—constitute the trilemma that the present paper sets out to navigate.

Prior research on parameter-efficient fine-tuning offers a rich palette of adaptation primitives, but these primitives have largely been investigated in static, text-based benchmarks or in image classification. LoRA injects trainable low-rank matrices into the linear projection layers of transformer blocks, effectively learning a residual path in a compact subspace [6]. Adapter layers insert small bottleneck modules between existing layers, with a down-

projection, a non-linearity, and an up-projection, thereby preserving the original feature space while learning task-specific offsets [7]. Prefix-tuning and its prompt-based relatives modify the input to the attention mechanism by prepending learnable continuous vectors, influencing the model's behavior without touching its core weights [8]. In video segmentation architectures, analogous strategies have been deployed, but they rarely incorporate runtime decision logic about which layers to adapt. A fixed LoRA rank, for instance, may be simultaneously too expressive for simple scenes, leading to overfitting and wasted computation, and insufficiently expressive for complex multi-object scenarios, resulting in persistent segmentation errors.

The notion of open-world dynamic scenes introduces additional demands that are not captured by standard fine-tuning benchmarks. Dynamic scenes are characterized by a continuous interplay of photometric variation, motion blur, partial occlusions, scene clutter, and the appearance of objects whose visual signatures were never encountered during training. UVO evaluates the ability of a model to segment all foreground objects irrespective of their semantic category, exposing the limitations of category-bound segmentation heads [10]. BURST enriches this setting with diverse annotation types and forces a unified treatment of recognition, segmentation, and tracking [11]. OVIS places particular emphasis on severe occlusion, where objects may disappear for many frames before re-emerging, challenging the memory's ability to retain and match object representations [12]. TAO broadens the evaluation to a federated corpus of mixed-domain videos, highlighting the brittleness of models that overfit to a single data distribution [13]. These benchmarks collectively demonstrate that the model's ability to handle openness is predicated not merely on the size of its pre-training corpus but on the adaptability of its internal routing and memory mechanisms.

When viewed from a systems perspective, the deployment of adaptive fine-tuning in such scenarios requires a feedback control loop that senses three classes of signals: the scene complexity signal, which condenses metrics such as the number of detected objects, the entropy of mask proposals, and the magnitude of optical flow; the model confidence signal, which encapsulates per-frame prediction uncertainty and temporal consistency metrics; and the resource availability signal, which reports the current GPU memory headroom, battery state, and network latency to a potential cloud offload target. An effective adaptive system must fuse these signals into a policy that determines the scope and intensity of parameter updates. Designing this policy without compromising the quality of the segmentation stream is the central architectural challenge addressed in the following section.

3. Architectural Design of an Adaptive Parameter Controller

The conceptual heart of an adaptive parameter-efficient fine-tuning system for video object segmentation is a meta-controller that negotiates between the granularity of adaptation modules and the momentary demands of the environment. We advocate for a hierarchical architecture organized into three tiers: a perception front-end that ingests video frames and produces initial segmentations using a frozen backbone, an adaptation backbone that hosts several parallel banks of injectable parameter-efficient modules, and a controller module that periodically decides, based on aggregated temporal statistics, which subset of modules will be updated and with what intensity. This architecture is deliberately separated into frozen and plastic components, echoing neurobiological principles of complementary learning systems, to minimize interference while allowing selective plasticity.

At the lowest tier, the perception front-end is instantiated by a large pre-trained video segmentation model such as SAM 2 with its accompanying memory encoder and mask

decoder. The front-end is kept entirely frozen during deployment to maintain the model's broad priors and to provide a stable reference signal against which the controller can measure drift. The output of the front-end consists of multi-scale mask predictions, feature embeddings for each segmented object, and a set of confidence scores. These outputs are passed not only to the downstream consumer but also to a scene analyzer subcomponent that maintains a sliding window of recent segmentation quality indicators, including the temporal Jaccard index between consecutive masks and the ratio of low-confidence regions.

The adaptation backbone houses a collection of parameter-efficient modules that can be dynamically gated. Within this collection, several design choices exist, each with distinct implications for the adaptation capacity–interference trade-off. Low-rank adaptors that target the query, key, and value projection matrices of the transformer's attention blocks provide a dense yet compact adaptation capacity distributed throughout the network depth. Adapter layers that sit at the output of each transformer block, on the other hand, offer a modular gradient pathway that often leads to more stable optimization when the data distribution shifts gradually, because the adapter's bottleneck naturally suppresses high-frequency noise. Prefix-based adaptation, which modifies the input latent state, can be particularly effective for encoding global scene attributes, such as illumination conditions or general motion patterns, without interfering with fine-grained object-level representations. A practical instantiation may hybridize these strategies, for instance by using prefix vectors to encode global illumination context while employing LoRA modules on the mask decoder to adjust for instance-specific shape variations.

The unique challenge of open-world video segmentation lies in the fact that the appropriate configuration of these modules is not known a priori and can change at the sub-second timescale. Consider a scenario where a surveillance camera observes a quiet street for several minutes, during which the scene analyzer reports consistently high confidence and low segmentation entropy. The controller may respond by disabling all adaptation modules, thus reducing power consumption and preventing unnecessary parameter drift. When a sudden incident introduces a collection of unfamiliar objects—a parade with costumes, for example—the confidence signals drop and the entropy of mask proposals spikes. The controller must then make a rapid yet informed decision: it might activate lightweight prefix vectors first to capture the global color and texture shift, and if the confidence does not recover, progressively engage deeper LoRA modules in the lower layers of the encoder. This staged escalation strategy acts as a form of neural resource management, prioritizing the cheapest and least intrusive adaptation before committing to more aggressive parameter updates.

A critical design parameter is the temporal granularity at which the controller re-evaluates its configuration. Updating parameters on every frame would consume too much compute and risk thrashing, where the model oscillates between different modes without converging. Conversely, updating only on long intervals of several seconds would miss transient events that are definitional to dynamic scenes. Recent work on memory-aware fine-tuning for SAM 2 has shown that it is possible to choreograph the interaction between the memory bank, which stores frame-level features, and the learnable parameters such that the system can handle sequences longer than those seen during pre-training without compromising memory efficiency [17]. This insight suggests a controller that operates at two distinct timescales: a fast timescale that updates only the memory encoding and attention biasing parameters via prefix vectors, and a slower timescale that periodically retrains the LoRA modules on a curated replay buffer of challenging frames identified by the scene analyzer. The interplay of

these two timescales can be viewed as a hierarchical temporal abstraction, where the fast loop maintains temporal coherence and the slow loop gradually adjusts the model's inductive bias.

The architectural decisions outlined above are subject to a set of structural trade-offs that must be understood in detail. The most immediate trade-off is between adaptation capacity and inference latency. Engaging additional LoRA modules increases the total number of multiply–accumulate operations per frame, which could push the inference time beyond the inter-frame interval on resource-constrained hardware. A scheduler that is hardware-aware can dynamically modulate the rank of LoRA modules or the width of adapter bottlenecks in accordance with a monitored latency budget. For instance, when operating on an embedded GPU with limited tensor core throughput, the controller might offload part of the adaptation computation to a nearby edge server via a split-computation paradigm. This introduces a second trade-off involving network latency and privacy. Offloading intermediate features or gradient information to the edge server can enable richer adaptation through more powerful compute, but it exposes visual information that may contain personally identifiable content. System designers must therefore build into the controller a privacy budget that restricts offloading when sensitive objects, such as faces or license plates, are detected, a topic to which we return in the governance section.

Memory footprint constitutes the third axis of this trade-off space. The SAM 2 architecture already maintains a substantial memory bank of spatiotemporal embeddings, and augmenting it with multiple sets of trainable adapter parameters can exhaust the limited high-bandwidth memory of edge accelerators. To mitigate this, the controller can implement a strategy of selective forgetting, in which adapter weights that have not contributed to improving segmentation confidence over a defined horizon are pruned or quantized. This biological metaphor of synaptic pruning is directly applicable to parameter-efficient modules because their small size makes it feasible to maintain multiple candidate weights and select the most informative subset. Such strategies elevate the controller from a mere decision module to a full-fledged resource manager that orchestrates the entire life cycle of adaptation.

4. Infrastructure, Deployment, and Sustainability Considerations

Bringing an adaptive parameter-efficient fine-tuning system from a laboratory prototype to a sustained, large-scale deployment requires careful attention to the surrounding computational and organizational infrastructure. In contemporary practice, video object segmentation systems are increasingly deployed across a continuum of compute tiers, ranging from battery-operated IoT cameras to regional cloud data centers. The heterogeneity of this hardware ecosystem introduces a significant deployment challenge because the same policy that optimally balances accuracy and latency on a high-end server GPU may be entirely infeasible on a low-power drone platform. Therefore, the adaptive controller architecture described above must be situated within a broader orchestration layer that handles model distribution, versioning, telemetry collection, and energy accounting.

A central material concern is energy sustainability. The rapid proliferation of deep learning models in always-on video analytics has raised alarms about the aggregate carbon footprint of training and inference. Fine-tuning a large foundation model on a new video domain, even when using parameter-efficient methods, still incurs non-trivial energy costs, especially if the adaptation occurs frequently across thousands of camera streams. Researchers in the green AI community have established frameworks for reporting energy consumption alongside accuracy metrics, and these frameworks advocate for a shift from accuracy-at-all-costs to a multi-objective optimization that penalizes excessive energy use [20]. When embedded into

the adaptive controller, sustainability becomes a first-class design constraint. The controller can be equipped with a carbon-intensity monitor that receives signals from a grid energy dashboard, and it can voluntarily throttle the frequency or depth of adaptation updates during periods of high carbon intensity. Although this might result in temporarily higher segmentation errors, the aggregate societal benefit of reducing emissions can be substantial when aggregated across a global camera fleet.

The deployment architecture also intersects strongly with data governance and privacy regulations. Video streams captured in public spaces are subject to an evolving patchwork of legal frameworks, such as the General Data Protection Regulation in Europe and various municipal ordinances elsewhere, that impose constraints on the storage, processing, and cross-border transfer of visual data. An adaptive fine-tuning system that continuously extracts frames for replay buffers or that offloads feature tensors to central servers must comply with data minimization principles. This necessitates a governance-by-design approach in which the controller's data handling procedures are auditable and configurable by region. For instance, a deployment spanning multiple countries could compile a geo-policy file that specifies which adaptation modules may be activated in each jurisdiction and whether cloud offloading is permitted. Such a file would be interpreted by the controller at runtime to dynamically reconfigure its privacy posture, a capability that mirrors software-defined networking but for machine learning pipelines.

Federated learning offers a complementary paradigm for distributed adaptation that preserves data locality while still allowing model improvement across a fleet of cameras. In a federated configuration, each edge node performs parameter-efficient fine-tuning on its local video stream and periodically sends only the modified adapter weights or LoRA matrices to a central aggregation server, where they are merged and redistributed [21]. This approach dramatically reduces the communication overhead compared with transferring raw video frames and, crucially, aligns with data sovereignty requirements because sensitive visual information never leaves the device. However, federated aggregation of parameter-efficient modules raises its own set of robustness concerns, including the risk of model poisoning by malfunctioning or malicious nodes and the challenge of reconciling adapters trained on non-independent and non-identically distributed data streams. A production-grade adaptive framework must therefore integrate anomaly detection at the federation level, monitoring statistical signatures of submitted weight updates and quarantining nodes whose gradients deviate excessively from the cohort median.

Continuous integration and delivery of model updates constitute another infrastructural pillar that is often underestimated in research settings. In a deployment that spans thousands of instances, it is unrealistic to assume that all nodes will run the same version of the base segmentation model or the same set of adaptation policies. A robust deployment pipeline must support staggered rollouts, automated canary testing on a fraction of the fleet, and the ability to rapidly roll back updates that degrade segmentation quality on under-represented scene types. Moreover, the telemetry data collected from the fleet—observational data such as mean confidence scores, adaptation trigger rates, and hardware utilization—serve as a vital feedback signal to the design team, enabling iterative refinement of the controller's hyperparameters. The design of this telemetry system must balance the richness of the diagnostic signals with the overhead of data transmission and the imperative to avoid capturing sensitive information. Techniques such as on-device feature extraction that sends

only aggregate statistics, along with differential privacy mechanisms that inject calibrated noise into these statistics, can help reconcile these competing pressures.

5. Robustness, Fairness, and Failure Mode Analysis in Dynamic Environments

The move from static to dynamic open-world settings fundamentally alters the robustness profile of video object segmentation systems. In a static, closed-world benchmark, robustness is typically measured by the model's ability to withstand synthetic perturbations such as Gaussian noise or compression artifacts. In open-world dynamic scenes, however, robustness must be conceived more broadly to encompass the model's resilience to distributional shifts, its capacity to detect and recover from its own failure modes, and its behavior on long-tailed subpopulations that were under-represented in training data. Adaptive fine-tuning, while offering a mechanism to improve performance on novel data, also introduces new fragility pathways. When the controller continuously updates parameters based on recent observations, it risks latching onto spurious correlations that hold transiently in a particular scene but generalize poorly to other scenes or even to future frames of the same sequence. For example, a segmentation model adapting to a foggy morning might unlearn its ability to segment distant vehicles under clear conditions, thereby degrading safety in the afternoon.

Understanding and cataloging failure modes is therefore a prerequisite for safe deployment. A taxonomic analysis reveals at least four classes of failure that are unique to adaptive parameter-efficient systems. The first is runaway adaptation, where a series of low-confidence frames triggers an aggressive parameter update that dramatically alters the segmentation mask of a critical object, leading to a cascade of tracking failures. The second is catastrophic interference between adaptation modules that target overlapping representational spaces, causing the model to oscillate between two incompatible sets of weights. The third is resource starvation, where the controller misestimates the computational cost of an adaptation step and exhausts the available memory or compute budget, leaving downstream tasks without timely segmentation results. The fourth is fairness degradation, where the adaptive process inadvertently amplifies biases present in the data stream. For instance, if the camera predominantly captures pedestrians of a particular demographic group during the adaptation window, the adapted model might become more accurate on that group while exhibiting blunders on others, effectively creating a discriminatory segmentation system. Researchers in sociotechnical fairness have warned that data-driven abstractions in machine learning pipelines can mask such inequities and that continuous monitoring of fairness metrics is essential [19].

Mitigating these failure modes calls for a multilayered defense-in-depth strategy. At the architectural layer, the controller can impose bounds on the magnitude of parameter updates per unit time, a form of trust-region constraint that limits the extent to which a single batch of frames can reshape the model. Additionally, a shadow model, which remains frozen, can be run in parallel and its output compared with that of the adapted model. When the discrepancy between the two exceeds a pre-calibrated threshold, the controller can roll back the recent adaptation steps. This shadow-model watchdog serves a purpose analogous to safety monitors in autonomous vehicles and provides a tractable mechanism for auditing the model online. At the data layer, the replay buffer used for adaptation can be curated to maintain stratified samples across object categories, illuminations, and motion types, thereby reducing the risk of catastrophic forgetting and bias amplification.

Fairness must be operationalized through quantitative metrics that are continuously tracked during deployment. Metrics such as the per-category mask intersection-over-union across

demographic segments, where such segments can be inferred from visual attributes or auxiliary metadata, provide a running indicator of whether the adaptive system is deepening or narrowing representational disparities. When a fairness metric crosses a predefined bound, the controller can adjust its adaptation policy, for instance by down-weighting frames from over-represented subpopulations in the replay buffer or by temporarily freezing adaptation until the metric recovers. This notion of fairness-aware adaptive control integrates insights from the algorithmic fairness literature with the unique temporal dynamics of video segmentation, opening new research questions about the interplay between fairness and online continual learning.

6. Governance, Policy, and Socio-Technical Implications

The deployment of adaptive parameter-efficient fine-tuning in open-world video object segmentation is not merely a technical endeavor but a socio-technical practice that is embedded within institutional, legal, and ethical frameworks. The capacity of these systems to silently reconfigure their internal parameters based on locally observed data introduces an opaque layer of behavior that challenges existing accountability structures. When a segmentation failure contributes to a consequential decision—such as a collision by an autonomous vehicle or a false identification in a law enforcement context—it becomes extremely difficult to attribute responsibility because the model state that produced the faulty output may have been shaped by an accumulation of decentralized adaptation steps across multiple nodes and time scales. Establishing a chain of custody for model updates that is as robust as the software supply chain for critical infrastructure is thus an urgent governance requirement.

One approach to institutionalizing accountability involves embedding cryptographic provenance into the adaptation pipeline. Each parameter update, whether it originates from a local replay buffer or a federated aggregation round, can be signed and logged in an immutable tamper-evident ledger. While this does not guarantee correctness, it provides a forensic capability to trace faults back to specific adaptation events and to identify the environmental conditions that triggered them. Such an infrastructure would need to be complemented by regulatory frameworks that mandate regular audits of adaptive AI systems, akin to the financial audits imposed on publicly traded companies. The European Union’s proposed AI Act, with its categorization of high-risk AI systems, already establishes a trajectory toward mandatory conformity assessments, and adaptive video segmentation systems deployed in safety-critical contexts would likely fall under such provisions.

The privacy implications of continual adaptation are similarly profound. Even when raw video frames are not stored, the fact that the model’s parameters are modified in response to individual-level data can theoretically enable membership inference attacks that reveal whether a particular person was present during adaptation. The integration of formal privacy guarantees, such as differentially private stochastic gradient descent applied to parameter-efficient modules, must therefore be considered a baseline requirement rather than an optional enhancement. Furthermore, the adaptive controller’s decisions about which data to retain in its replay buffer constitute a form of data curation that may be subject to data protection laws. System operators will need to build interfaces that allow data subjects to request the deletion of their influence, a right enshrined in GDPR, which translates, in our context, to rolling back adaptation steps that relied on data traces linked to that individual. The technical feasibility of such machine unlearning for parameter-efficient adapters is an open and critical research area.

Beyond privacy and accountability, the governance of adaptive segmentation systems must engage with the broader societal question of surveillance. Open-world video segmentation is increasingly deployed in public safety, retail analytics, and smart city applications, where its ability to identify and track any moving object—regardless of its semantic category—creates a novel form of ambient monitoring. The fact that the system adapts to new objects on the fly means that its surveillance capability can expand over time in ways that are not anticipated during procurement. Policy interventions are needed to ensure that the scope of surveillance remains aligned with publicly deliberated values. These interventions might include mandatory transparency reports that disclose the categories of objects a system has learned to segment, impact assessments conducted prior to deployment in sensitive areas, and sunset clauses that automatically disable adaptation after a predefined operational period unless explicitly renewed. By embedding these mechanisms into the procurement and operations lifecycle, municipalities and organizations can reap the benefits of adaptive visual intelligence while safeguarding civil liberties.

7. Conclusion

This paper has presented a system-level analysis of adaptive parameter-efficient fine-tuning for open-world video object segmentation in dynamic scenes, arguing that the field must move beyond accuracy-centric model design toward holistic architectures that sense, schedule, audit, and govern their own adaptation. We have delineated a conceptual architecture in which a meta-controller orchestrates a portfolio of parameter-efficient modules—low-rank adaptors, adapter layers, and prefix vectors—guided by real-time signals of scene complexity, model confidence, and resource availability. The structural trade-offs among latency, memory consumption, and energy sustainability were examined, revealing the need for hardware-aware scheduling and staged escalation strategies. Deployment considerations ranging from edge-cloud partitioning to federated learning and green AI were integrated into the design space, underscoring that sustainability and privacy are not afterthoughts but formative constraints. Robustness and fairness were analyzed as emergent properties that must be continuously monitored through a defense-in-depth strategy that includes shadow models, trust-region constraints, and stratified replay buffers. Finally, the paper situated these systems within the wider socio-political context, arguing for cryptographic provenance, regulatory audits, privacy-preserving learning, and transparent governance mechanisms to prevent mission creep in surveillance capabilities.

Looking forward, the evolution of adaptive parameter-efficient fine-tuning will be propelled by advances in meta-learning, automated machine learning, and system-on-chip memory architectures, but its societal integration will be shaped by legal, ethical, and participatory processes. Interdisciplinary research that combines video understanding, systems engineering, law, and science and technology studies is therefore essential. By embracing this broader framing, the community can build segmentation infrastructures that are not only accurate and efficient but also robust, fair, and accountable under the infinite variability of the open world.

References

1. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., & Girshick, R. (2023). Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision, 4015–4026.

2. Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Radle, R., Rolland, C., Gustafson, L., & Feichtenhofer, C. (2024). SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714.

3. Cheng, H. K., & Schwing, A. G. (2022). XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In European Conference on Computer Vision (pp. 640–658). Springer.

4. Yang, Z., Wei, Y., & Yang, Y. (2021). Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems, 34, 2491–2502.

5. Yang, Z., & Yang, Y. (2022). Decoupling features in hierarchical propagation for video object segmentation. Advances in Neural Information Processing Systems, 35, 25202–25214.

6. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.

7. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning (pp. 2790–2799). PMLR.

8. Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (pp. 4582–4597). Association for Computational Linguistics.

9. Ghiasi, G., Gu, X., Cui, Y., & Lin, T.-Y. (2022). Scaling open-vocabulary image segmentation with image-level labels. In European Conference on Computer Vision (pp. 540–557). Springer.

10. Wang, W., Feiszli, M., Wang, H., & Tran, D. (2021). Unidentified video objects: A benchmark for dense, open-world segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10776–10785).

11. Athar, A., Luiten, J., Hermans, A., Ramanan, D., & Leibe, B. (2023). BURST: A benchmark for unifying object recognition, segmentation and tracking in video. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1674–1683).

12. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P. H. S., & Bai, S. (2022). Occluded video instance segmentation: A benchmark. International Journal of Computer Vision, 130(8), 2022–2039.

13. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., & Ramanan, D. (2020). TAO: A large-scale benchmark for tracking any object. In European Conference on Computer Vision (pp. 436–454). Springer.

14. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1290–1299).

15. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

16. Yang, L., Fan, Y., & Xu, N. (2019). Video instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5188–5197).

17. Li, G., Yuan, H., Chen, S., Hu, Q., Wang, J., & Jiang, K. (2026). MFT: Memory-Aware Fine-Tuning of SAM2 for Efficient Long-Sequence Video Object Segmentation. IEEE Signal Processing Letters.

18. Karimi Mahabadi, R., Henderson, J., & Ruder, S. (2021). Compacter: Efficient low-rank hypercomplex adapter layers. In Advances in Neural Information Processing Systems, 34, 1022–1035.

19. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68). ACM.

20. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

21. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

22. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54–71.