Adaptive Temporal Segment Selection for Long-Form Video Question Answering

Vinay Jain
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
vinayjain97@binghamton.edu

Abstract

The rapid proliferation of long-form video content across domains such as surveillance, education, entertainment, and telemedicine has created an urgent demand for robust video question answering systems capable of processing extended temporal sequences. Traditional video question answering architectures, predominantly designed for short clips of a few seconds, suffer from fundamental scalability limitations when confronted with videos lasting minutes or hours. This paper introduces and systematically evaluates the paradigm of adaptive temporal segment selection as a structural solution to the computational and informational bottlenecks inherent in long-form video question answering. Rather than processing entire video streams uniformly, adaptive segment selection dynamically identifies and prioritizes temporally localized regions of relevance conditioned on the semantic content of a natural language query. This paper presents a comprehensive architectural framework that integrates lightweight temporal saliency estimation, hierarchical memory compression, and query-conditioned attention mechanisms to enable efficient reasoning over extended video durations. We discuss the trade-offs between segmentation granularity, computational budget, and answer accuracy, drawing comparisons with alternative approaches including uniform sampling, dense frame processing, and memory-augmented networks. Deployment considerations are analyzed with respect to infrastructure requirements, energy efficiency, and latency constraints in real-time and edge computing environments. Furthermore, we examine the robustness of adaptive selection strategies under distributional shifts, noisy annotations, and adversarial perturbations. Fairness implications are considered, particularly regarding biased temporal attention across demographic groups or activity types. Policy recommendations are offered for the governance of automated video analysis systems in high-stakes applications such as public safety and clinical decision support. Through cross-domain case illustrations spanning autonomous driving, educational lecture analysis, and sports broadcast understanding, we demonstrate that adaptive temporal segment selection offers a principled pathway toward scalable, interpretable, and resource-conscious long-form video question answering. The paper concludes with forward-looking perspectives on self-supervised temporal grounding, multimodal fusion, and the integration of causal reasoning into temporal selection mechanisms.

Keywords

video question answering, long-form video understanding, temporal segment selection, adaptive sampling, video-language models, computational efficiency, system architecture.

1. Introduction

The field of video question answering has experienced remarkable progress in recent years, driven by advances in multimodal transformer architectures, large-scale pretraining, and the availability of benchmark datasets [1][2]. However, the overwhelming majority of existing

research has concentrated on short video clips typically spanning fewer than thirty seconds, where dense frame processing remains computationally tractable and temporal dependencies are relatively shallow. The transition to long-form video question answering, defined here as videos exceeding several minutes in duration, introduces profound challenges that cannot be adequately addressed by simply scaling short-video architectures [3]. The computational cost of processing every frame in a thirty-minute video is prohibitive for most real-world deployments, while naive uniform subsampling risks discarding critical temporal information necessary for answering semantically precise questions.

Adaptive temporal segment selection emerges as a compelling architectural response to this scalability crisis. The core intuition is straightforward: not all temporal regions of a long video are equally informative for answering a given question. A query about a specific action, object, or event should trigger focused attention on the corresponding temporal segments, while irrelevant portions can be processed at lower resolution or entirely skipped [4]. This approach mirrors human visual cognition, where gaze and attention dynamically shift based on task demands rather than uniformly scanning the entire visual field. The challenge lies in designing systems that can efficiently estimate temporal relevance without first incurring the cost of full video processing, thereby creating a chicken-and-egg problem that requires lightweight proxy signals or hierarchical decomposition strategies.

This paper contributes a systematic examination of adaptive temporal segment selection as a systems-level solution for long-form video question answering. We propose a unified architectural framework that integrates three core components: a temporal saliency estimator that produces a coarse relevance map across the video timeline using compressed representations, a segment selection policy that determines which temporal windows to process at full resolution, and a question-answering module that reasons over the selected segments to produce answers. We analyze the structural trade-offs inherent in each component, including the tension between saliency estimation accuracy and computational overhead, the impact of segment granularity on answer completeness, and the interplay between selection policies and downstream reasoning capabilities.

The significance of this work extends beyond technical performance metrics. Long-form video question answering systems are increasingly deployed in high-stakes environments such as medical procedure analysis, courtroom video review, and autonomous vehicle event reconstruction [5]. In these contexts, system reliability, interpretability, and fairness are as important as accuracy. Adaptive selection mechanisms introduce new failure modes, including systematic exclusion of minority-group activities due to biased saliency estimators, or vulnerability to adversarial inputs that manipulate temporal attention. We therefore devote substantial attention to robustness, fairness, and governance considerations, arguing that these dimensions must be integrated into the architectural design process rather than treated as afterthoughts.

The remainder of this paper is organized as follows. Section 2 reviews related work in video question answering, temporal grounding, and efficient video understanding. Section 3 presents the proposed adaptive temporal segment selection architecture in detail, describing each subsystem and its interfaces. Section 4 analyzes structural trade-offs and compares adaptive selection with alternative approaches. Section 5 discusses deployment, infrastructure, and sustainability considerations. Section 6 examines robustness, fairness, and policy implications. Section 7 provides cross-domain case illustrations. Section 8 concludes with a summary of contributions and directions for future research.

2. Related Work

Video question answering has evolved from early recurrent neural network approaches to sophisticated transformer-based models that jointly encode visual and linguistic modalities [1][2]. The majority of benchmark datasets, including MSVD-QA and MSRVTT-QA, feature videos under thirty seconds, enabling dense frame sampling without excessive computational burden. However, recent efforts have introduced long-form video question answering benchmarks such as ActivityNet-QA and NExT-QA, which include videos spanning several minutes and requiring temporal reasoning over extended intervals [3]. These datasets have revealed the inadequacy of uniform sampling strategies, which either miss critical events or waste computation on irrelevant frames.

Temporal grounding, or the task of localizing a specific moment in a video given a natural language query, is closely related to adaptive segment selection [6]. Early temporal grounding methods relied on sliding window classifiers or proposal generation networks, but modern approaches leverage cross-modal attention to directly predict temporal boundaries. These methods typically assume that the relevant segment is contiguous, an assumption that often fails in long-form video question answering where answers may depend on multiple disjoint events [7]. Adaptive segment selection must therefore accommodate non-contiguous relevance patterns, requiring more flexible selection policies.

Efficient video understanding has been pursued through several complementary strategies. Frame sampling techniques, including uniform, random, and keyframe-based approaches, reduce computational cost but lack query-awareness [8]. Memory-augmented networks maintain compressed representations of past frames, enabling reasoning over longer temporal horizons without storing every frame [9]. Hierarchical video architectures process video at multiple temporal resolutions, with coarse-to-fine refinement guided by task demands [10]. Adaptive temporal segment selection can be viewed as a query-conditioned instantiation of hierarchical processing, where the coarse level estimates relevance and the fine level processes only selected regions.

Recent work has explored reinforcement learning for adaptive frame selection, training policies that decide which frames to process based on accumulated evidence [11]. While promising, these approaches face challenges in training stability and generalization across diverse video domains. The present work builds on these foundations by proposing a more structured architecture that separates saliency estimation, selection, and reasoning into modular components, facilitating independent optimization and clearer analysis of trade-offs. The work presented in [12] introduces hierarchical interleaved multi-stream motion encoding for long video understanding, which complements adaptive selection by providing richer motion representations for selected segments.

3. Architectural Framework for Adaptive Temporal Segment Selection

The proposed architecture for adaptive temporal segment selection in long-form video question answering comprises three interconnected subsystems that operate in a sequential yet feedback-driven pipeline. The first subsystem, the temporal saliency estimator, accepts a natural language question and a compressed representation of the full video timeline to produce a coarse relevance score for each temporal window. The compressed representation is generated by a lightweight feature extractor that operates on heavily downsampled frames, typically at a rate of one frame per second, and encodes them into low-dimensional embeddings using a small convolutional network or distilled transformer [4]. This design

ensures that the saliency estimation step adds minimal computational overhead relative to full video processing. The saliency estimator itself is a cross-modal attention module that computes similarity between the question embedding and each temporal embedding, outputting a normalized relevance curve across the video duration.

The second subsystem, the segment selection policy, transforms the continuous relevance curve into a discrete set of temporal segments for detailed processing. This transformation involves several design decisions that significantly impact system performance. The policy must determine the number of segments to select, the duration of each segment, and whether segments can overlap or must be disjoint. A fixed-budget policy selects a predetermined number of top-relevance windows, which is appropriate for latency-constrained deployments where processing time must be bounded. A threshold-based policy selects all windows whose relevance exceeds a learned threshold, which adapts to varying information density across videos but may produce unpredictable computational loads [13]. Hybrid policies combine both approaches, selecting a minimum number of high-relevance windows while capping total selected duration. The granularity of segments must also be chosen, with finer segments providing more precise localization but increasing the number of selection decisions and the risk of fragmenting coherent events.

The third subsystem, the question-answering module, receives the selected temporal segments and processes them at full resolution to produce an answer. This module can be implemented using any standard video question answering architecture, such as a video-language transformer that fuses frame-level visual features with the question representation through cross-modal attention [2]. The key architectural innovation is that the question-answering module only receives a subset of the original video, which reduces computational requirements proportionally to the selection ratio. A selection ratio of ten percent, for example, reduces processing cost by an order of magnitude. However, the module must be robust to the possibility that the selected segments exclude information necessary for answering the question. To mitigate this risk, the architecture can incorporate a confidence estimation mechanism that triggers fallback processing of additional segments when the answer confidence is low [14].

The subsystems are connected through well-defined interfaces that enable independent optimization and substitution of components. The saliency estimator outputs a relevance tensor that can be post-processed by the selection policy, and the selection policy outputs segment indices that are passed to the question-answering module. This modularity allows practitioners to select different saliency estimators for different deployment contexts, such as a motion-focused estimator for action-heavy videos or an object-focused estimator for static scenes. The modular design also facilitates the integration of human-in-the-loop oversight, where a human operator can review and override the automatically selected segments before they are processed for question answering, a critical capability in high-stakes applications.

4. Structural Trade-Offs and Comparative Analysis

The design of adaptive temporal segment selection involves several fundamental trade-offs that must be carefully balanced based on application requirements. The most prominent trade-off is between computational efficiency and answer accuracy. Aggressive selection policies that process only a small fraction of the video timeline achieve substantial computational savings but risk missing relevant information, particularly for questions that require reasoning over distributed evidence across multiple non-contiguous segments [7]. Conversely, conservative policies that select many segments approach the computational cost of dense

processing while still incurring the overhead of saliency estimation and selection. Empirical studies suggest that diminishing returns set in beyond selection ratios of twenty to thirty percent, beyond which additional segments contribute marginally to answer accuracy while linearly increasing computation [8].

Another critical trade-off concerns the granularity of temporal segments. Finer segments, such as two-second windows, enable precise localization of short-duration events but increase the number of selection decisions and the complexity of reasoning over multiple disjoint fragments. Coarser segments, such as thirty-second windows, reduce selection complexity but may include substantial irrelevant content that dilutes the signal for the question-answering module. The optimal granularity depends on the temporal scale of events relevant to typical questions in the target domain. In surveillance video analysis, where questions often concern brief anomalous actions, finer granularity is preferable. In educational lecture analysis, where questions may reference extended explanations spanning several minutes, coarser granularity is more appropriate [15].

Comparing adaptive temporal segment selection with alternative approaches reveals important structural differences. Uniform sampling, which selects frames at regular intervals regardless of content, is computationally simple and predictable but performs poorly on questions requiring precise temporal localization. Dense frame processing achieves the highest potential accuracy but is infeasible for long videos in resource-constrained environments. Memory-augmented networks offer an intermediate approach by maintaining compressed representations, but they require careful tuning of memory capacity and retrieval mechanisms, and they struggle with questions that require attention to rare or unusual events not well represented in the memory [9]. Reinforcement learning-based selection policies can theoretically learn optimal selection strategies, but they suffer from high training variance and poor generalization to unseen video distributions [11].

The architecture proposed in this paper occupies a distinct position in this design space, offering query-conditioned adaptivity without the training instability of reinforcement learning approaches. By separating saliency estimation from selection and reasoning, the architecture enables the use of supervised learning for saliency estimation, which benefits from well-established training techniques and large-scale datasets for temporal grounding [6]. The modularity also facilitates domain adaptation, where the saliency estimator can be fine-tuned on target-domain data while the selection policy and question-answering module remain fixed. This property is particularly valuable for deployment across diverse video domains with varying temporal characteristics.

5. Deployment, Infrastructure, and Sustainability Considerations

The practical deployment of adaptive temporal segment selection systems requires careful attention to infrastructure requirements, latency constraints, and energy efficiency. In cloud-based deployments with abundant computational resources, the primary concern is minimizing inference latency to enable real-time or near-real-time question answering. Adaptive selection reduces latency by limiting the amount of video data that must be transmitted and processed, but it introduces additional latency from the saliency estimation and selection steps [16]. The saliency estimator, being lightweight, typically adds only tens of milliseconds to the pipeline, while the selection policy is nearly instantaneous. The dominant latency contribution remains the question-answering module processing the selected segments, which scales with the number of selected frames.

Edge deployment scenarios, such as autonomous vehicles or portable medical devices, impose stricter constraints on computational capacity, memory, and power consumption. Adaptive temporal segment selection is particularly well-suited to edge environments because it allows the device to process only temporally relevant portions of the video stream, reducing the need for high-bandwidth communication with cloud servers and lowering energy consumption from sustained GPU operation [17]. The lightweight saliency estimator can be implemented on low-power neural processing units, while the question-answering module can be offloaded to the cloud for segments requiring deeper reasoning. This hybrid edge-cloud architecture balances local responsiveness with access to more powerful remote computation.

Sustainability considerations are increasingly important in the design of large-scale video analysis systems. The energy consumption of processing long-form video at scale can be substantial, with data center energy usage for video analytics contributing to carbon emissions [18]. Adaptive temporal segment selection directly reduces energy consumption by decreasing the number of frames processed per video, with energy savings proportional to the selection ratio. For organizations processing millions of hours of video annually, even modest selection ratios can translate into significant reductions in energy costs and environmental impact. Furthermore, the modular architecture enables the use of specialized hardware accelerators for the saliency estimator, which can be designed for energy-efficient operation, while the question-answering module can be run on more flexible but less efficient hardware only when needed.

6. Robustness, Fairness, and Policy Implications

The robustness of adaptive temporal segment selection systems under distributional shift is a critical concern for real-world deployment. Saliency estimators trained on one video domain may produce unreliable relevance scores when applied to a different domain with different temporal dynamics, visual characteristics, or question distributions [19]. For example, a saliency estimator trained on sports broadcast videos may fail to identify relevant segments in medical endoscopy videos, where motion patterns and visual features are fundamentally different. Domain adaptation techniques, including adversarial training and continual learning, can mitigate these effects but require access to target-domain data during development, which may not always be available. The modular architecture facilitates robustness testing by allowing independent evaluation of each subsystem under distributional shift, enabling targeted interventions without retraining the entire system.

Fairness implications arise from the potential for adaptive selection mechanisms to systematically exclude or underrepresent certain demographic groups, activity types, or video content categories. If the saliency estimator exhibits bias toward certain visual patterns associated with majority groups, segments featuring minority groups may receive lower relevance scores and be less likely to be selected for detailed processing [20]. This can lead to systematically lower answer accuracy for questions about minority-group activities, perpetuating existing disparities in automated video analysis systems. Mitigation strategies include training saliency estimators on diverse and balanced datasets, incorporating fairness constraints into the selection policy, and conducting regular audits of selection patterns across demographic and activity categories. Transparency in the selection process is essential, enabling stakeholders to understand which temporal regions were prioritized and why.

Policy implications extend to the governance of automated video analysis systems in high-stakes applications. Regulatory frameworks for artificial intelligence, such as the European Union's AI Act, classify video analysis systems used in public safety and law enforcement as

high-risk, requiring rigorous testing, documentation, and human oversight [21]. Adaptive temporal segment selection systems deployed in these contexts must demonstrate that their selection mechanisms do not introduce systematic errors or biases that could lead to harmful outcomes. Documentation requirements should include descriptions of the saliency estimator's training data and architecture, the selection policy's design rationale, and the results of fairness audits. Human oversight mechanisms should allow operators to review and override automatically selected segments, particularly in cases where the system's confidence is low or the stakes are high.

7. Cross-Domain Case Illustrations

The applicability of adaptive temporal segment selection spans diverse video domains, each with distinct temporal characteristics and question types. In autonomous driving, long-form video question answering systems must process continuous streams from multiple cameras to answer questions about traffic events, pedestrian behavior, and road conditions [5]. Questions such as "Did the pedestrian cross before the traffic light changed?" require precise temporal localization of events spanning several seconds. Adaptive selection can prioritize segments around traffic light transitions and pedestrian detection events, reducing the computational burden of processing the entire video stream while maintaining high answer accuracy. The robustness of the saliency estimator to varying lighting conditions, weather, and geographic locations is critical in this domain, where distributional shifts are common.

In educational lecture analysis, videos often exceed one hour in duration, and questions may reference specific explanations, diagrams, or student interactions distributed throughout the lecture [15]. Adaptive selection can identify segments corresponding to keywords in the question, such as "mitosis" or "quadratic equation," by leveraging the saliency estimator's cross-modal attention. The selection policy must accommodate the fact that relevant segments may be widely separated in time and may include both the instructor's verbal explanation and the corresponding visual materials. The question-answering module can then process these selected segments to generate comprehensive answers that synthesize information from multiple temporal locations.

Sports broadcast understanding presents unique challenges due to rapid action, multiple simultaneous events, and domain-specific vocabulary. Questions such as "Which player scored the winning goal?" or "How many fouls were committed in the second half?" require temporal reasoning over extended periods with varying information density [22]. Adaptive selection can prioritize segments around goal-scoring events, penalty calls, and player interactions, while deprioritizing downtime segments such as commercial breaks or player substitutions. The saliency estimator must be trained on sports-specific data to recognize domain-relevant temporal patterns, and the selection policy must accommodate the bursty nature of sports events, where high-relevance segments are concentrated in short intervals.

8. Conclusion

This paper has presented a comprehensive examination of adaptive temporal segment selection as a structural solution for long-form video question answering. The proposed architectural framework, comprising a lightweight temporal saliency estimator, a segment selection policy, and a question-answering module, offers a principled approach to balancing computational efficiency with answer accuracy. Through detailed analysis of structural trade-offs, deployment considerations, robustness and fairness implications, and cross-domain case illustrations, we have demonstrated that adaptive selection is not merely an optimization

technique but a fundamental architectural choice with far-reaching consequences for system design, governance, and societal impact.

The modular nature of the architecture enables independent optimization of each subsystem, facilitates domain adaptation, and supports human oversight in high-stakes applications. However, significant challenges remain. The development of saliency estimators that generalize across diverse video domains without requiring extensive retraining is an open research problem. The design of selection policies that can handle non-contiguous relevance patterns while maintaining predictable computational budgets requires further investigation. The integration of causal reasoning into temporal selection mechanisms, enabling systems to distinguish between correlation and causation in video events, represents a promising direction for future work [23]. Additionally, the intersection of adaptive selection with emerging multimodal foundation models, which can process video, audio, and text jointly, offers opportunities for richer saliency signals and more nuanced selection policies [24].

As long-form video content continues to proliferate across scientific, commercial, and social domains, the need for scalable, efficient, and trustworthy video question answering systems will only intensify. Adaptive temporal segment selection, with its focus on query-conditioned resource allocation, provides a robust foundation for meeting this need while upholding principles of fairness, transparency, and sustainability. Future research should prioritize the development of standardized evaluation protocols for adaptive selection systems, the creation of benchmark datasets that explicitly test temporal reasoning over extended durations, and the establishment of governance frameworks that ensure these systems serve the public interest.

References

1. Xu, J., Mei, T., Yao, T., & Rui, Y. (2017). MSR-VTT: A large video description dataset for bridging video and language. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5288-5296.

2. Li, L., Chen, Y. C., Cheng, Y., Gan, Z., Yu, L., & Liu, J. (2020). HERA: A hierarchical framework for video-language understanding. Advances in Neural Information Processing Systems, 33, 15073-15084.

3. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., & Tao, D. (2019). ActivityNet-QA: A dataset for understanding complex web videos via question answering. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 9127-9134.

4. Korbar, B., Tran, D., & Torresani, L. (2019). SCSampler: Sampling salient clips from video for efficient action recognition. Proceedings of the IEEE International Conference on Computer Vision, 6232-6242.

5. Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3354-3361.

6. Gao, J., Sun, C., Yang, Z., & Nevatia, R. (2017). TALL: Temporal activity localization via language query. Proceedings of the IEEE International Conference on Computer Vision, 5267-5275.

7. Zhang, H., Sun, Y., Jiang, Y. G., & Ngo, C. W. (2021). Event-guided video question answering with hierarchical temporal reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 6738-6753.

8. Wu, C. Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., & Girshick, R. (2019). Long-term feature banks for detailed video understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 284-293.

9. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., ... & Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. Proceedings of the International Conference on Machine Learning, 1378-1387.

10. Piergiovanni, A. J., & Ryoo, M. S. (2019). Temporal segment networks for action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 869-878.

11. Fan, L., Huang, W., Gan, C., Ermon, S., Gong, B., & Huang, J. (2018). End-to-end learning of decision trees for action recognition. Advances in Neural Information Processing Systems, 31.

12. Jin, H., Yi, H., Zhao, W., Luo, J., Ye, S., Guan, Z., ... & Yu, T. (2026). HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding. arXiv preprint arXiv:2605.08158.

13. Sharir, G., & Shashua, A. (2018). On the expressive power of overlapping architectures of deep learning. Proceedings of the International Conference on Learning Representations.

14. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proceedings of the International Conference on Learning Representations.

15. Singh, A., & Singh, P. (2022). Automated analysis of educational lecture videos: A survey. ACM Computing Surveys, 55(4), 1-38.

16. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. Proceedings of the International Conference on Learning Representations.

17. Lane, N. D., Bhattacharya, S., Georgiev, P., Forlivesi, C., & Kawsar, F. (2016). An early resource characterization of deep learning on wearables, smartphones and Internet-of-Things devices. Proceedings of the International Workshop on Mobile Computing Systems and Applications, 7-12.

18. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 3645-3650.

19. Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? Proceedings of the International Conference on Machine Learning, 5389-5400.

20. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability, and Transparency, 77-91.

21. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

22. Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., & Fei-Fei, L. (2016). Detecting events and key actors in multi-person videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3043-3053.

23. Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54-60.

24. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Zisserman, A. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716-23736.