Inquiry: The University of Arkansas Undergraduate Research Journal

Volume 23 Article 5

July 2024

Survey of Memory Consolidation Techniques for Video Question Answering

Matthew Couts
University of Arkansas, Fayetteville

Pha Nguyen
University of Arkansas, Fayetteville

Khoa Luu
University of Arkansas, Fayetteville

Survey of Memory Consolidation Techniques for Video Question Answering

Cover Page Footnote

This research has been supported by an Honors College Research Grant.

Survey of Memory Consolidation Techniques for Video Question Answering

Matt Couts (mrcouts@uark.edu)

Pha Nguyen (panguyen@uark.edu)

Faculty Mentor - Khoa Luu (khoaluu@uark.edu)

Department of Electrical Engineering and Computer Science, College of Engineering, University of Arkansas, Fayetteville, AR 72758

Submitted: May 6, 2024

Accepted: June 23, 2024

Abstract

Video Question Answering (VideoQA) is a field of research focused on developing models that can engage in natural conversations with humans about the content of videos. Currently, the most successful approaches involve analyzing videos frame-by-frame, which is computationally and memory-intensive. To imitate human memory, the Atkinson-Shiffrin memory model can formulate the machine's video understanding capability through Vision-Language Models. Reducing the number of frames processed by the model is a crucial operation in this approach category and can be handled by a memory consolidation algorithm. The memory consolidation algorithm should be able to determine the keyframes to transfer from short-term to long-term memory. However, due to the complexity of events in videos, this approach may need to pay more attention to critical information by efficient and appropriate operations. This paper aims to compare video understanding capabilities by analyzing the memory consolidation

algorithms. Specifically, we present experiments evaluating simple but effective memory consolidation operations on the ActivityNet-QA dataset to construct an optimal memory consolidation process.

Introduction

Large Language Model-based assistants, i.e., chatbots, like the well-known ChatGPT and GPT-4 (Achiam et al., 2023), exhibit a remarkable ability to answer a wide range of questions posed to them. This capacity has sparked interest in replicating that proficiency in multimodal applications (Li, Li, Savarese, & Hoi, 2023). This interest has resulted in the development of visual question-answering programs, which aim to answer questions related to images. Interestingly, visual question-answering programs have demonstrated notable success in their domain. In fact, OpenAI recently began rolling out visual question-answering understanding features to ChatGPT (OpenAI, 2023).

Programs that look to extend large language models (LLMs) to other modalities are known as Multimodal Large Language Models (MLLMs). A popular approach to developing vision-language MLLMs involves aligning frozen, pre-trained vision and language models (Li et al., 2023; Maaz, Rasheed, Khan, & Khan, 2023). It avoids the high training costs associated with training vision-language models. Cross-modal alignment, then, is the focus of this approach. BLIP-2 (Li, Li, Savarese, & Hoi, 2023) achieves cross-modal alignment using a lightweight transformer, e.g., Q-Former, that serves as the middleman between the frozen LLM and the frozen image encoder, providing the LLM with the most useful visual features. Q-Former has served as an integral piece in video understanding programs (Zhang, Li, & Bing, 2023). While this approach has improved the state-of-the-are in video understanding tasks, these models struggle to handle long videos (Song et al., 2023; Zhang, Li, & Bing, 2023). It is because these models rely on frame-wise feature extraction. Thus, long videos with thousands of frames computationally overwhelm these models.

The Atkinson-Shiffrin model (Atkinson & Shiffrin, 1968) is proposed as a viable approach to improving long video understanding (Song et al., 2023). It is structured around a short-term and long-term memory buffer, as illustrated in Fig. 1. The model utilizes a sliding window to analyze the video. It shows that the Atkinson-Shiffrin model reduces both computational complexity and memory costs and improves long-term temporal connection (Song et al., 2023). Our analysis identifies potential structural limitations that may hinder the model's performance.

In this study, we investigate several simple operational modifications to the Atkinson-Shiffrin model, specifically exploring variations of the memory consolidation algorithm. Our objective is to observe and evaluate the impact of these modifications on the overall performance of the VideoQA tasks.

Related Work

Multi-modal Large Language Models

Given the recent success of Large Language Models (LLMs), it is only natural to look to equip these models with multi-modal capabilities. BLIP-2 (Li, Li, Savarese, & Hoi, 2023) presents a notable framework for designing MLLMs built around a Query transformer (Q-Former) that links a frozen pre-trained image encoder with a frozen pre-trained LLM. The BLIP-2 framework has become a popular approach to building MLLMs. Video-LLaMA (Zhang, Li, &

Bing, 2023) creates and utilizes a video Q-Former and an audio Q-Former to build an MLLM with both image, video, and audio understanding abilities.

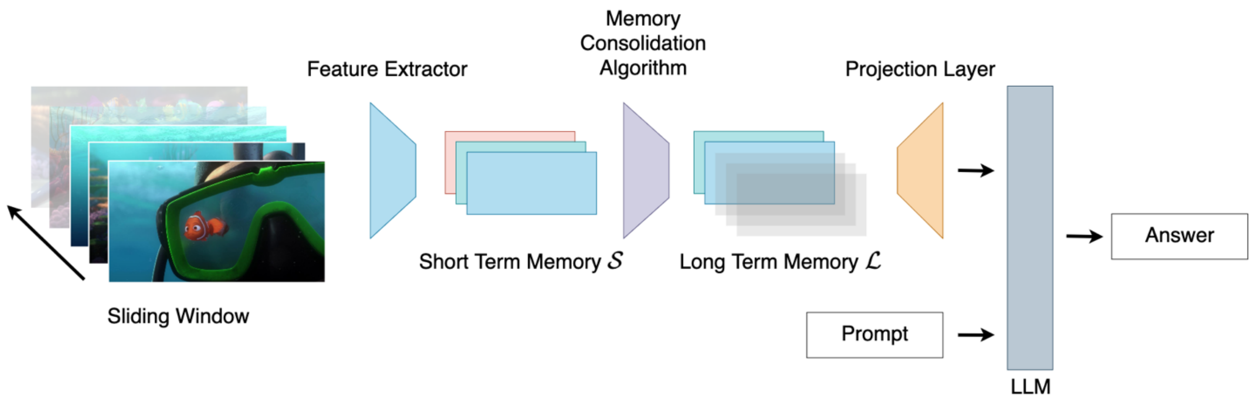

Figure 1 The Atkinson-Shiffrin model is proposed for the VideoQA task. It is structured around short-term and long-term memory.

Long Video Understanding

Handling long videos is a challenge facing video understanding frameworks (Song et al., 2023; Zhang, Li, & Bing, 2023). Considering the use of image-based models for frame-wise feature extraction, the struggle of these frameworks in understanding long videos is unsurprising. As videos get longer, the number of frames grows, and memory requirements grow exponentially. MovieChat (Song et al., 2023) is a recent work that uses the Atkinson-Shiffrin memory model to reduce the number of frames that need to be handled by the program.

The number of frames is reduced through a hierarchical clustering memory consolidation algorithm that merges pairs of consecutive frames based on the pair's similarity. This approach is built on the assumption that temporally adjacent frames within a movie scene, for example, are typically redundant.

Methodology

Problem Setup

Video Question Answering (VideoQA) is the task of responding to natural language prompts about the contents of a video. A VideoQA program takes a video and a prompt as inputs and should output a response to the prompt in natural language that is correct, given the context of the input video. A program that can perform VideoQA well must have a solid understanding of the text, audio, and video modalities and how the modalities align with one another.

This paper focuses on the text and video modalities specifically. LLMs have shown to handle the text modality well (Achiam et al., 2023; Touvron et al., 2023; Touvron et al., 2023) and are a critical part of most VideoQA programs. Video foundation models have improved at representing the contents of a video as tokens, but struggle to capture long-term temporal connections in long videos. The struggle of video foundation models to represent videos accurately has limited the video understanding of VideoQA models. This research explores the use of a memory consolidation algorithm as a method of improving the ability of VideoQA to videos.

Overall Framework

The framework includes a visual feature extractor, a short-term and long-term memory buffer, and a large language model. The system takes a video and a prompt as inputs. First, the input video is divided into smaller video clips. Each video clip is then converted on a frame-by-frame basis into a set of abstract numerical representations, called tokens, by way of the visual feature extractor. These tokens are packed with the information that is used to develop a response to the input prompt.

Keeping a large number of frames in memory concurrently is very expensive. To avoid this, the framework processes the video using a sliding window, which breaks the video into several fragments. The sliding window allows the visual feature extractor to process the video without needing to store all of its frames at once. Though the sliding window alleviates memory requirements related to feature extraction, storing the tokens produced by the feature extractor remains a concern.

To address this, this framework is structured around the Atkinson-Shiffrin memory model. The idea is that, following generation, tokens will be moved into short-term memory. In short-term memory, a memory consolidation algorithm will consolidate the tokens into a smaller set of representative tokens. These representative tokens are merged into long-term memory. The long-term representative tokens are passed to the LLM to respond the input prompt.

Visual Feature Extractor

For visual feature extraction, the framework avoids using video-based foundation models like ViViT (Arnab et al., 2021). Instead, we use an image-based model to extract features from the video on a frame-by-frame basis. The decision to forgo the use of video-based foundation models revolves around the lack of a video foundation model that aligns well with the text modality. This framework uses EVA-CLIP (Fang et al., 2023) and the Q-Former from the BLIP-2 (Li, Li, Savarese, & Hoi, 2023) for image feature extraction. Video clips formed by the sliding window are passed into the visual feature extractor, frame-by-frame.

Memory Consolidation

The frame features held in the short-term memory buffer are merged into representative frames by the memory consolidation algorithm. These representative frames are stored in the long-term memory and presented to the LLM for response generation. The memory consolidation algorithm was built with the assumption that consecutive frames are often redundant and that videos can be understood with the set of most informative frames. Thus, the memory consolidation algorithm is built to reduce redundancy.

Given a set of frame features, the memory consolidation algorithm calculates a metric between pairs of consecutive frame features. The pair with the highest score is merged and remains in short-term memory. Again, the metric between pairs of consecutive frame features is calculated, and the pair with the highest score is merged. This process repeats until only two frames remain. The remaining two frames are moved to long-term memory and the short-term memory buffer is cleared, ready for the frame features of the following video clip. An illustration of the baseline memory consolidation algorithm can

be found in Fig. 2. Below are the different memory consolidation techniques we explored in this research.

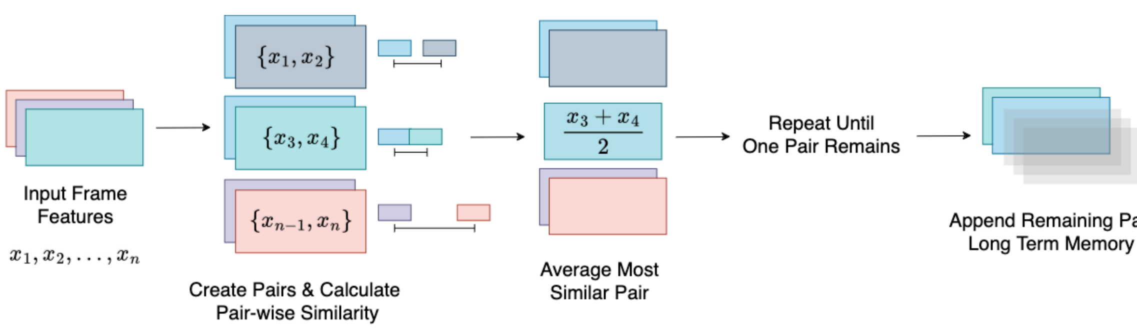

Figure 2 The core memory consolidation algorithm.

No Consolidation. To gain the impact the memory consolidation algorithm has on the model's ability to perform well on the VideoQA task, we removed the memory consolidation algorithm altogether. Removing the consolidation algorithm allows us to compare the value in the Atkinson-Shiffrin memory model and determine whether the added model complexity is worth it.

Cosine Similarity. This approach is shown in Alg. 1. The assumption behind this approach is that movies often contain scenes that are slowed down to emphasize a moment, but for video understanding, the movie can be understood without

redundancy.

Algorithm 1 Similarity-based Consolidation Algorithm

Require: $\mathcal{S}$ ▷ short-term memory
while $len(\mathcal{S}) > R_L$ **do** ▷ iterative merge
  for $\mathbf{x}_i$ in $\mathcal{S}$ **do**
    $s \leftarrow cosine\_sim(\mathbf{x}_i, \mathbf{x}_{i+1})$ ▷ tokens similarity
  end for
  $m \leftarrow max(s)$ ▷ the maximum value index
  $\mathbf{x}_m \leftarrow average(\mathbf{x}_m, \mathbf{x}_{m+1})$ ▷ merge
  del $\mathbf{x}_{m+1}$
end while

Cosine Distance. Instead of merging frames based on what consecutive frame feature pair is most similar, we merge frames based on what consecutive frame pair is most different. To do this, we take the frame feature pair with the highest cosine distance instead of the highest cosine similarity. It was done assuming that this would improve the model's ability to detect anomalies in videos. This approach is similar to Alg. 1, with similarity replaced by distance.

Furthest Frame Pair. The goal of the memory consolidation algorithm is to merge the frames from the short-term memory buffer into representative frames that best capture the ideas presented by the video clip. Taking the two frames that were furthest from each other, without requiring them to be consecutive, would present the two most distinct frame features from the video clip, as presented in Alg. 2.

Algorithm 2 Furthest Pair Consolidation Algorithm

Require: $\mathcal{S}$ ▷ short-term memory
  for $\mathbf{x}_i$ in $\mathcal{S}$ **do**
    for $\mathbf{x}_j$ in $\mathcal{S}$ where $j > i$ **do**
      $s \leftarrow cosine\_dist(\mathbf{x}_i, \mathbf{x}_j)$
      if $s > max$ **then** ▷ max distance
        $max = [x_i, x_j]$
      end if
    end for
  end for
  $\mathcal{L}.append(max)$

Large Language Model

The baseline model uses 7 billion parameters Llama 2 (Touvron et al., 2023). More specifically, they use the Vicuna-7b-1.5, an LLM created by fine-tuning Llama 2. The LLM is frozen in this model and aligned with the video modality by a Q-Former as presented in BLIP-2 (Li, Li, Savarese, & Hoi, 2023). We make no changes to the LLM used by the baseline model in our survey.

Experimental Details

Dataset

To assess the VideoQA abilities of the modified model, we utilize the ActivityNet-QA dataset (Yu et al., 2019), a fully annotated, large-scale dataset for VideoQA. We evaluate the model's zero-shot VideoQA capabilities via the evaluation process as the same as in Video-ChatGPT (Maaz, Rasheed, Khan, & Khan, 2023). We also use this dataset to collect some qualitative evaluations. Additionally, we use the UCF-Crime dataset (Sultani, Chen, & Shah, 2018), which consists of untrimmed surveillance videos containing anomalies from 13 anomaly categories. This dataset does not support the VideoQA task, to evaluate the model's ability to detect anomalies in videos.

Evaluation Metrics

The ActivityNet-QA dataset contains three types of questions, including motion, spatial relationship, and temporal relationship (Yu et al., 2019). The dataset conducts evaluation based on accuracy but expects only a "yes" or "no" response. Our model, however, produces human-like, free-form textual responses and elaborates further than a "yes" or "no". Thus, to test the accuracy of our model, we use GPT to compare our model's responses to the ground truth as in other work (Maaz, Rasheed, Khan, & Khan, 2023). GPT scores are based on its subjective understanding of the question, the ground truth, and our model's response. It means that the evaluation scores will vary on a run-to-run basis, as shown in Table 1, which is undesirable. This aspect of our evaluation of ActivityNet-QA is important to keep in mind. We note this gap in VideoQA evaluation methods as a topic that needs further exploration.

Table 1. Evaluation Results Differ Run-to-Run

Run	Yes Count	No Count	Accuracy	Score
1	85	355	0.193	1.718
2	95	345	0.216	1.875
3	89	351	0.202	1.850

Quantitative Results

Each modification to the model is quantitatively assessed using a subset of the ActivityNet-QA dataset. The accuracy metric represents the percentage of yes or no questions that the model answered correctly. Since our model generates more elaborate responses that do not explicitly state "yes" or "no," we need a way to infer whether the model's response implies a yes or a no to the input prompt. To achieve this, we used GPT to interpret our model's response as a yes or a no and used this interpretation to determine the model's accuracy. The score metric, provided by GPT, measures the semantic relevance of the model's response to the input prompt. Of the two metrics, we believe that answering questions correctly is more critical in practical applications of VideoQA models. For that reason, our interpretation of the results values accuracy more than it does score.

Of the methods we tested, the cosine similarity memory consolidation algorithm had the highest accuracy, outperforming the next best algorithm, the furthest frame pair, by 10%. The algorithm with the highest score was the furthest frame pair with a score of 3.00, followed closely by the cosine similarity with a score of 2.97. Results can be found in Table 2. Despite its lower score compared to the furthest frame pair algorithm, we conclude cosine similarity as the best operation for VideoQA models due to its significantly higher accuracy and nearly identical score. Our conclusion is also built upon our assumption that accuracy should be weighted more heavily than score when evaluating models for VideoQA.

Table 2. Evaluation results on Activitynet-QA dataset

Modification	**Accuracy**	**Score**
No Consolidation	0.290	2.22
Furthest Frame Pair	0.305	**3.00**

Cosine Distance	0.275	2.37
Cosine Similarity	**0.405**	2.97

Qualitative Results

We gather qualitative results using both the ActivityNet-QA dataset to evaluate the general VideoQA capabilities of the model and the UFC-Crime dataset to gauge the anomaly detection abilities of the model. To evaluate the general VideoQA capabilities of the model, we prompt the model with generalized questions as shown in Fig. 3. In evaluating the anomaly detection abilities of the model, we tailor our questions to the anomalous event present in the video, as shown in Fig. 4.

The results from our qualitative analysis were promising. The responses from the Atkinson-Shiffrin memory model contained accurate information, for the most part. However, there was a problematic number of hallucinations or inaccurate information, leaving plenty of room for improvement in the model's capabilities.

Q: What is happening in this video?

A: The video shows a man cutting a man's hair in a salon. The man is cutting his hair while a woman watches him. The salon is equipped with a hairdryer, curtains, a brick wall, a black shirt, a black dress, a white wall, a black and white image, and a black purse.

Figure 3 Question and answer concerning a video from the ActivityNet-QA dataset.

Text in green is accurate information, text in yellow is unrelated or unclear information, and text in red is inaccurate information.

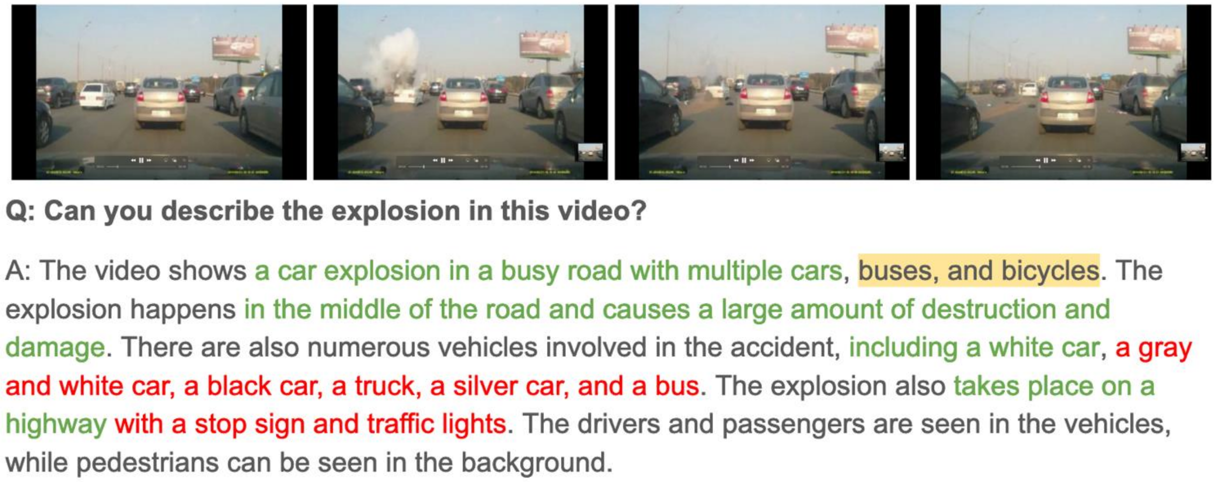

Figure 4 Question and answer concerning a video from the UCF-Crime dataset (Sultani, Chen, & Shah, 2018).

Text in green is accurate information, text in yellow is unrelated or unclear information, and text in red is inaccurate information.

Conclusion

In this work, we have demonstrated the potential of the Atkinson-Shiffrin memory model to enhance machine video comprehension when paired with an effective memory consolidation algorithm.

In our quantitative study, the cosine similarity memory consolidation algorithm performed the best. However, we note that the metrics used, despite being standard among similar research, are undesirable as results rely on the subjective analysis of a large language model. We note this as an opportunity for future research to explore better, standardized evaluations.

In our qualitative analysis, the models demonstrated good video understanding capabilities, but still included a troubling number of hallucinations. We hypothesize that the model's tendency to elaborate unnecessarily as an avenue for inaccuracies to be introduced in the model's response but leave further exploration for another time.

Vision-Language Models have become a topic of interest in machine video understanding. Our survey of memory consolidation algorithms builds upon previous research by identifying the best performing consolidation strategy for the Atkinson-Shiffrin memory model.

Acknowledgements

This research has been supported by an Honors College Research Grant.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6836-6846).

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89-195). Academic press.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., ... & Cao, Y. (2023). Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19358-19369).

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., ... & Qiao, Y. (2023). Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Maaz, M., Rasheed, H., Khan, S., & Khan, F. S. (2023). Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

OpenAI, G. P. T. (2023). 4V (ision) System Card. *preprint*.

Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., ... & Wang, G. (2024). Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18221-18232).

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6479-6488).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., & Tao, D. (2019, July). Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 9127-9134).

Zhang, H., Li, X., & Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

About the Authors

Matt Couts

Matt Couts is an honors undergraduate student studying computer science at the University of Arkansas. In Fall of 2023, Matt joined the Computer Vision and Image Understanding (CVIU) Lab at the university, where he has conducted undergraduate research in image and video understanding systems. He was awarded an Honors College Research Grant to support his undergraduate research for the Spring and Fall 2024 semesters.

Pha Nguyen

Pha Nguyen is a Ph.D Candidate in Computer Science in Computer Vision and Image Understanding (CVIU) Lab, University of Arkansas. Before pursuing his doctorate study, he was a Research Engineer at VinAI Research Institute and obtained his B.Sc degree from Vietnam National University - University of Science. His research interests are focused on computer vision and its application in image and video understanding systems, such as Visual Temporal Modeling, Vision-Language Models, Multiple Object Tracking. He has served as a reviewer for premier conferences and journals, including NeurIPS, CVPR, ECCV, ICCV, CVIU, and IEEE Access.

Khoa Luu

Dr. Luu is an Assistant Professor and the Director of the Computer Vision and Image Understanding (CVIU) Lab in the Department of Electrical Engineering and Computer Science (EECS) at the University of Arkansas (UA), Fayetteville, US. He is also affiliated with the Center for Public Health and Technology, UA, and the NSF MonARK Quantum Foundry. He is serving as an Associate Editor of the IEEE Access Journal and the Multimedia Tools and Applications Journal, Springer Nature. He also serves as the Area Chair in the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 and 2024, and the Conference on Neural Information Processing Systems (NeurIPS) 2024.

He was the Research Project Director at the Cylab Biometrics Center at Carnegie Mellon University (CMU), USA. His research interests focus on various topics, including Smart Health, Precision Agriculture, Quantum Machine Learning, Multi-Object Tracking, Human Behavior Understanding, Scene Understanding, Face Recognition, Domain Adaptation, Image and Video Processing, Deep Learning and Compressed Sensing. He has received six patents and three Best Paper Awards and coauthored 120+ papers in conferences, technical reports, and journals.