Anil Kag

Senior Research Scientist | Lead, Efficient Generative AI
Snap Research

Los Angeles, CA 90066
+1 (347) 489 6990
anilkagak2@gmail.com
anilkagak2.github.io
Citations: 1100+ | h-index: 15

Education

2018–2023 **Ph.D. in Electrical & Computer Engineering**, *Boston University*, MA, USA, *GPA: 3.96 / 4.0*
- **Dissertation:** Novel Neural Architectures & Algorithms for Efficient Inference
- **Committee:** Profs. Venkatesh Saligrama (Advisor), Brian Kulis (BU), Alexander Olshevsky (BU), Kilian Q. Weinberger (Cornell), and Prateek Jain (Google DeepMind).

2018–2022 **M.S. in Electrical & Computer Engineering**, *Boston University*, MA, USA, *GPA: 3.96 / 4.0*

2010–2014 **B.Tech. in Computer Science**, *Indian Institute of Technology*, Guwahati, India, *GPA: 9.20 / 10*
- **Ranked 4th out of 80** students in the Computer Science Class of 2014.

Work Experience

April 2025–Present **Senior Research Scientist (L5) | Team Lead, Efficient GenAI**, *Snap Research*, Santa Monica, CA

Lead **Efficient Generative AI Team**, directing the technical strategy for Snap's hardware-agnostic generative stack.
- **Elastic Supernetworks:** Spearheading the development of **SnapGen++**, a pioneering Elastic Diffusion Transformer (DiT) framework. This "supernetwork" enables a single training run to deploy optimized models across disparate hardware (iPhone, Android, Server), delivering competitive fidelity to models 10x larger, such as **Flux**.
- **Efficiency Frontiers:** Architecting the next generation of **SnapVideo** models, achieving superior generation quality at significantly lower inference compute than state-of-the-art public models like **WAN**.
- **Mobile Video Pioneers:** Directing the team in the design of high-performance DiT models capable of **real-time streaming video generation on-device**, a landmark achievement in mobile computational photography.
- **Strategic Mentorship:** Leading a team of research scientists to bridge the gap between foundational research and production-scale AI deployment.

July 2023–March 2025 **Research Scientist (L4)**, *Snap Research*, Santa Monica, CA

Developed foundational generative architectures that transitioned from academic highlights to core Snapchat product features.
- **SnapGen (Patent-Pending):** Authored the landmark **SnapGen** paper (**CVPR 2024 Highlight**); orchestrated the architecture for which **Snap Inc. filed a United States Patent**, recognizing the unique novelty of the mobile image generation model.
- **Large-Scale Product Impact:** Engineered **SnapVideo v2**, delivering end-to-end training optimizations for large-scale video generation. This model currently **powers numerous AI Lenses** used by hundreds of millions of Snapchat users daily.
- **Scholarly Leadership:** Mentored over 10 Research Interns, resulting in a prolific output of state-of-the-art papers at top-tier conferences (CVPR, NeurIPS) focusing on **Diffusion Models, RLHF Alignment, and Quantization**.

2016–2018 **Research Fellow**, *Microsoft Research*, Bangalore, India

Advanced the field of Extreme Multi-label Classification (XMC) for industrial recommendation systems.
- Developed **Parabel** and **SwiftXML**, core algorithms that achieved a **13.6% increase in CTR** for Bing Ads, impacting multi-million dollar annual revenue streams.

2014–2016 **Software Engineer**, *Microsoft*, Bangalore, India

Optimized the Interactive Service Hub for Dynamics CRM, reducing page load times by over 50%.

Patents & Intellectual Property

Summary Author of **9 United States Patent Applications** filed by **Snap Inc.** covering foundational inventions in Efficient Diffusion Transformers and Generative Media Synthesis.

2025
- **ELITE**: One Model, Many Budgets: Elastic Latent Interfaces for Diffusion Transformers.
  Patent Status: Approved for Filing (Sept 2025) | Associated Paper
- **Sprint**: Sparse-Dense Residual Fusion for Efficient Diffusion Transformers.
  Patent Status: Approved for Filing (Sept 2025) | Associated Paper
- **DenseDPO**: Fine-Grained Temporal Preference Optimization for Video Diffusion.
  Patent Status: Approved for Filing (May 2025) | Associated Paper

2024
- **SnapGen**: Taming High-Resolution Text-to-Image Models for Mobile Devices.
  Patent Status: US Patent Filed (Nov 2024) | Associated Paper
- **BitsFusion**: 1.99 bits Weight Quantization of Diffusion Model.
  Patent Status: US Patent Filed (May 2024) | Associated Paper
- **Sf-V**: Single Forward Video Generation Model.
  Patent Status: US Patent Filed (May 2024) | Associated Paper

2023
- **Mind the Time**: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis.
  Patent Status: US Patent Filed (Nov 2023) | Associated Paper
- **AsCAN**: Asymmetric Convolution-Attention Neural Networks.
  Patent Status: US Patent Filed (Nov 2023) | Associated Paper
- **TextCraft**: Your Text Encoder Can be Image Quality Controller.
  Patent Status: US Patent Filed (Nov 2023) | Associated Paper

Thematic Research & Publications

Aggregated Citations: 1,100+ | h-index: 15 | Top ML Conference Papers: 20+ (Source: Google Scholar, March 2026)

Foundational Image Gen — Architecting efficient, high-resolution diffusion models for edge and large-scale deployment.
- **arXiv'26: SnapGen++**: Unleashing Diffusion Transformers for Efficient High-Fidelity Image Generation on Edge.
  D. Hu, A. Gupta, M. Gabidolla, A. Sahni, H. Coskun, Y. Li, Y. Idelbayev, A. Mahmood, A. Lebedev, D. Lahiri, A. Goyal, J. Hu, M. Gong, S. Tulyakov, **A. Kag**
- **ICLR'26: SPRINT**: Sparse-Dense Residual Fusion for Efficient Diffusion Transformers.
  D. Park, M. Haji-Ali, Y. Li, W. Menapace, S. Tulyakov, H. Kim, A. Siarohin, **A. Kag**
- **CVPR'26: Omni-Attribute**: Open-vocabulary Attribute Encoder for Visual Concept Personalization.
  T. Chen, A. Siarohin, G. G. Qian, K. C. J. Wang, E. Nemchinov, M. Haji-Ali, R. A. Guler, W. Menapace, I. Skorokhodov, **A. Kag**, J. Zhu, S. Tulyakov
- **CVPR'25 (Highlight): SnapGen**: Taming High-Res T2I Models for Mobile with Efficient Architectures.
  J. Chen, D. Hu, X. Huang, H. Coskun, A. Sahni, A. Gupta, A. Goyal, D. Lahiri, R. Singh, Y. Idelbayev, J. Cao, Y. Li, K. Cheng, S. Chan, M. Gong, S. Tulyakov, Y. Xu, J. Ren, **A. Kag**
- **NeurIPS'24: BitsFusion**: 1.99 bits Weight Quantization of Diffusion Model.
  Y. Sui, Y. Li, **A. Kag**, Y. Idelbayev, J. Cao, J. Hu, D. Sagar, B. Yuan, S. Tulyakov, J. Ren
- **NeurIPS'24: AsCAN**: Asymmetric Convolution-Attention Networks for Efficient Recognition and Generation.
  A. Kag, H. Coskun, J. Chen, J. Cao, W. Menapace, A. Siarohin, S. Tulyakov, J. Ren

Foundational Video Gen — Pioneering high-fidelity, spatiotemporal transformers for mobile and server-side synthesis.
- **CVPR'26: S2DiT**: Sandwich Diffusion Transformer for Mobile Streaming Video Generation.
  L. Zhao, Y. Wu, A. Lebedev, D. Lahiri, M. Dong, A. Sahni, M. Vasilkovsky, H. Chen, J. Hu, A. Siarohin, S. Tulyakov, Y. Wang, **A. Kag**, Y. Li
- **CVPR'26: ELITE**: One Model, Many Budgets: Elastic Latent Interfaces for Diffusion Transformers.
  M. Haji-Ali, W. Menapace, I. Skorokhodov, D. Park, **A. Kag**, M. Vasilkovsky, S. Tulyakov, V. Ordonez, A. Siarohin
- **CVPR'25: SnapGen-V**: Generating a Five-Second Video within Five Seconds on a Mobile Device.
  Y. Wu, Z. Zhang, Y. Li, Y. Xu, **A. Kag**, Y. Sui, H. Coskun, K. Ma, A. Lebedev, J. Hu, D. Metaxas, Y. Wang, S. Tulyakov, J. Ren
- **CVPR'24 (Highlight): Snap Video**: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis.
  W. Menapace, A. Siarohin, I. Skorokhodov, E. Deyneka, T. S. Chen, **A. Kag**, Y. Fang, A. Stoliar, E. Ricci, J. Ren, S. Tulyakov
- **NeurIPS'24: Sf-V**: Single Forward Video Generation Model.
  Z. Zhang, Y. Li, Y. Wu, Y. Xu, **A. Kag**, I. Skorokhodov, W. Menapace, A. Siarohin, J. Cao, D. Metaxas, S. Tulyakov, J. Ren
- **NeurIPS'25: PointVid**: Towards Physical Understanding in Video Generation: A 3D Point Regularization Approach.
  Y. Chen, J. Cao, V. Goel, S. Korolev, C. Jiang, J. Ren, S. Tulyakov, **A. Kag**
- **arXiv'25: H3AE**: High Compression, High Speed, and High Quality AutoEncoder for Video Diffusion Models.
  Y. Wu, Y. Li, I. Skorokhodov, **A. Kag**, W. Menapace, S. Girish, A. Siarohin, Y. Wang, S. Tulyakov
- **arXiv'25: SnapGen-V2**: Taming Diffusion Transformer for Real-Time Mobile Video Generation.
  Y. Wu, Y. Li, **A. Kag**, I. Skorokhodov, W. Menapace, K. Ma, A. Sahni, J. Hu, A. Siarohin, D. Sagar, Y. Wang, S. Tulyakov

RLHF & Alignment — Advancing preference optimization and reward flow frameworks for generative model fine-tuning.
- **arXiv'26: Diffusion-DRF**: Free, Rich, and Differentiable Reward for Video Diffusion Fine-Tuning.
  Y. Wang, Y. Li, G. G. Qian, S. Tulyakov, Y. Fu, **A. Kag**
- **NeurIPS'25 (Spotlight): DenseDPO**: Fine-Grained Temporal Preference Optimization for Video Diffusion Models.
  Z. Wu, **A. Kag**, I. Skorokhodov, W. Menapace, A. Mirzaei, I. Gilitschenski, S. Tulyakov, A. Siarohin
- **ICCV'25: RankDPO**: Scalable Ranked Preference Optimization for Text-to-Image Generation.
  S. Karthik, H. Coskun, Z. Akata, S. Tulyakov, J. Ren, **A. Kag**
- **CVPR'24: TextCraftor**: Your Text Encoder Can be Image Quality Controller.
  Y. Li, X. Liu, **A. Kag**, J. Hu, Y. Idelbayev, D. Sagar, Y. Wang, S. Tulyakov, J. Ren

Adaptive & Hardness-Aware	Developing intelligent systems that adapt model complexity based on input difficulty. ○ **ICLR'23: DiSK**: Scaffolding a Student to Instill Knowledge. **A. Kag**, D. A. E. Acar, A. Gangrade, V. Saligrama ○ **ICLR'23: Selective Query**: Efficient Edge Inference by Selective Query. **A. Kag**, I. Fedorov, A. Gangrade, P. Whatmough, V. Saligrama ○ **ICML'22 DyNN (Spotlight): TinyML**: Achieving High TinyML Accuracy through Selective Cloud Interactions. **A. Kag**, I. Fedorov, A. Gangrade, P. Whatmough, V. Saligrama ○ **NeurIPS'21 (Spotlight): Limited Feedback**: Online Selective Classification with Limited Feedback. A. Gangrade, **A. Kag**, A. Cutkosky, V. Saligrama ○ **AISTATS'21: OSP**: Learning With Abstention via One-Sided Classification. A. Gangrade, **A. Kag**, A. Cutkosky, V. Saligrama
Efficient RNN/CNN	Fundamental breakthroughs in low-complexity architectural design and optimization. ○ **CVPR'22: PDE-CNNs**: Condensing CNNs with Partial Differential Equations. **A. Kag**, V. Saligrama ○ **ICML'21: FPTT**: Training Recurrent Neural Networks via Forward Propagation Through Time. **A. Kag**, V. Saligrama ○ **CVPR'21: TARNN**: Time-Adaptive RNN: A Dynamical Systems View. **A. Kag**, V. Saligrama ○ **ICLR'20: iRNN**: RNNs Incrementally Evolving on an Equilibrium Manifold. **A. Kag**, Z. Zhang, V. Saligrama
Extreme Classification	Architecting industrial-scale recommendation systems for millions of labels. ○ **WWW'18: Parabel**: Partitioned Label Trees for Extreme Classification. Y. Prabhu, **A. Kag**, S. Harsola, R. Agrawal, M. Varma ○ **WSDM'18: SwiftXML**: Extreme Multi-label Learning with Label Features. Y. Prabhu, **A. Kag**, S. Gopinath, et al. ○ **NSDI'19: BLAS-on-flash**: An Efficient Alternative for Large Scale ML Training. S. J. Subramanya, H. V. Simhadri, S. Garg, **A. Kag**, V. Balasubramanian

Media & Press Coverage

Summary	Research and technological breakthroughs on **SnapGen** have been featured in major international trade publications and corporate newsroom announcements, highlighting the global impact of these mobile-first generative models.
2025	○ **TechCrunch**: Snap Unveils New AI Text-to-Image Model for Mobile Devices. Detailed coverage of the **SnapGen** architecture and its ability to deliver high-fidelity generation on-device. ○ **Snap Newsroom**: Snap Unveils Breakthrough AI Model Optimized for Mobile Devices. Official announcement regarding the innovation of **SnapGen** and its role in Snap's hardware-agnostic generative stack.

Academic Service

Peer Review	Evaluated **100+ research manuscripts** as a Program Committee Member or Invited Reviewer for top-tier Artificial Intelligence and Computer Vision venues.
Conferences	NeurIPS, CVPR, ICML, ICLR, ICCV, ECCV, AAAI, COLT, ICASSP
Journals	Transactions on Machine Learning Research (TMLR)
Recognition	Awarded **Top 10% Reviewer** status at **NeurIPS 2020** for exceptional service.

Academic & Professional Achievements

- **CVPR Highlight Award (2025 & 2024):** Research on *SnapGen* and *Snap Video* selected for Highlight presentation at the world's premier computer vision conference (Top 3–5% of global submissions).
- **NeurIPS Spotlight Award (2025 & 2021):** Recognized for high-impact contributions to Video Diffusion (*DenseDPO*) and Selective Classification (Top 3% of peer-reviewed papers).
- **ICML DyNN Workshop Spotlight (2022):** Selected for special oral presentation for pioneering work in *TinyML* and cloud-edge orchestration.
- **Top 10% Reviewer Recognition, NeurIPS (2020):** Formally recognized for "Extraordinary Service" in the peer-review process of the largest and most prestigious AI conference globally.
- **Rafik Hariri Graduate Student Fellowship:** Awarded to a select group of PhD candidates at Boston University for high-impact research potential in computational sciences.
- **Dean's Ph.D. Fellowship:** Meritorious multi-year fellowship provided by the ECE Department at Boston University.

- **IIT-G Merit-cum-Means Scholarship:** Awarded for sustained academic excellence during undergraduate studies at the Indian Institute of Technology (IIT).
- **IIT-JEE All India Rank 1761:** Secured a rank in the top **0.4% of 450,000+** candidates in one of the world's most competitive engineering entrance examinations.