Kevin Tian

kevinmtian@gmail.com | kevinmtian.github.io | linkedin.com/in/kevinmtian

Summary

Senior Machine Learning Engineer at ByteDance/TikTok (Singapore) and ex-Meta Research Scientist, building production AI systems across LLMs, multimodal foundation models, and generative vision. Focused on LLM post-training, agentic/RAG search, LLM-as-a-judge verification, video/audio/text pretraining, diffusion-based data synthesis, and end-to-end ML deployment.

Work Experience

ByteDance / TikTok *Singapore*

Senior Machine Learning Engineer *08/2023 – Present*

Focus: LLM Post-training & Agentic Search

- Architected a production-oriented LLM/SLM intent understanding system for Local Services Search, improving user-facing search relevance by converting ambiguous multilingual queries into structured POI, destination, brand, category, requirement, and intent signals.
- Built a unified compact model to act as extractor, query rewriter, POI recognizer, structured parser, LLM-as-a-judge verifier, and agent-style planner, reducing recall noise from inaccurate location/POI/intent parsing under resource-constrained serving settings.
- Improved structured understanding quality through LLM-assisted labeling, SFT/LoRA, and reinforcement-learning-based post-training, including PPO-style RLHF and DPO-style preference optimization, reaching strong precision/recall across multi-country evaluations.
- Designed nearline Kafka-to-cache deployment with canonicalized query rewriting, stable structured outputs, quantized multi-GPU inference, and high-throughput LLM serving, substantially improving cache coverage and reducing online LLM serving pressure while preserving latency/cost constraints.

Shenzhen University | PathoVision Co. Ltd. *Shenzhen, China*

Research Fellow | Technical Director *12/2020 - 6/2023*

Focus: Generative AI | Annotation-efficient 3D Medical Vision | Human-in-the-loop Learning

- Developed diffusion-based medical data synthesis methods to generate annotation-aligned image-mask pairs under limited-label settings, using mask-conditioned generation, texture-style injection, and frequency-domain attention to improve downstream segmentation robustness.
- Built weakly supervised segmentation frameworks from sparse scribble annotations, combining spatial self-attention, attentive similarity learning, partial segmentation loss, and masked CRF regularization to approach fully supervised performance on medical segmentation benchmarks.
- Developed interactive and online-learning 3D segmentation systems that converted sparse user inputs into proxy masks and used model residual maps to guide annotation, reducing expert labeling effort by 62% while maintaining competitive Dice performance.
- Led applied research and early-stage productization for privacy-sensitive healthcare AI workflows, translating generative data synthesis, weak supervision, continual learning, and human-in-the-loop 3D segmentation into deployable clinical prototypes.

Meta *Menlo Park, CA, USA*

Research Scientist *9/2017 - 11/2020*

Focus: Multimodal Foundation Models | Video/Audio/Text Understanding | ML Systems

- Developed multimodal video Transformer systems for large-scale harmful-content understanding, combining visual frames, audio/speech, transcripts, captions, metadata, and OCR-like signals to improve detection beyond single-modality models.
- Built self-supervised video/audio/text pretraining pipelines with cross-modal contrastive alignment and temporal sequence modeling, then fine-tuned models for supervised integrity classification across policy categories such as misinformation, adult content, violence, regulated goods, misleading ads, and normal videos.
- Designed label-wise attention pooling over encoded video sequences to produce both video-level policy predictions and temporal evidence, helping reviewers localize suspicious segments more efficiently instead of inspecting entire videos manually.
- Bridged research and production during Meta's Caffe2-to-PyTorch/TorchScript transition, reducing model-rewrite overhead and accelerating online evaluation; contributed to major reductions in political-misinformation prevalence and measurable improvements across other harmful-content categories.

Education

Ph.D., Applied Mathematics and Statistics, Stony Brook University, USA 2014–2017
M.S., Statistics, Tulane University, USA 2012–2014
B.S., Mathematics, East China Normal University, China 2008–2012

Skills

LLM post-training (SFT, LoRA, RLHF, PPO, DPO), diffusion models, multimodal transformers, agentic/RAG pipelines, LLM-as-a-judge, PyTorch, Python, C++, Spark, Hive, Flink, SQL, quantization, multi-GPU inference