# Shitong Shao (Sutton)

Generative AI Researcher – Ph.D. Student @ HKUST(GZ) – Video Diffusion, Distillation, and Efficient Generative Modeling

sshao213@connect.hkust-gz.edu.cn +86 15968506725 Homepage Google Scholar (770+ Citations) GitHub Douyin

## Technical Expertise

1. **Modeling and Distillation:** Step distillation (**MagicDistillation**, **FastLightGen** (**CVPR**), **AMD**); size distillation (**FastLightGen**); realtime distillation (realtime generation for **HYVideo-1.5** at **First-Intelligence**); VAE distillation (**HYVideo-1.5** VAE distillation with 4.5× acceleration); sparse attention (**PISA**, **LIVEditor**); dataset distillation (**G-VBSM** (**CVPR Highlight**), **EDC** (**NeurIPS**)); knowledge distillation (**TST** (**IJCAI Oral**), **Auto-DAS** (**ECCV**), **CKA** (**IJCAI Oral**)).
2. **Sampling and Inference Optimization:** Efficient sampling and few-step generation (**CoRe2** (**TPAMI**)); quality-oriented sampling and trajectory optimization (**IV-Mixed Sampler** (**ICLR**), **Golden Noise** (**ICCV**), **CoRe2** (**TPAMI**)).
3. **Application Domains:** Realtime video generation; text-to-image generation; text-to-video generation; text-image-audio-to-video generation; unified video editing.

## Research Profile

1. **Efficient generative modeling and video diffusion distillation.** My recent work centers on efficient video diffusion modeling across multiple operational scales, including **step**, **model size**, **realtime inference**, and **VAE compression**. Representative projects include **MagicDistillation**, **FastLightGen**, **AMD**, **PISA**, and **LIVEditor**.
2. **Sampling optimization and generation quality.** I study how efficient sampling, initial-noise design, and trajectory control can jointly improve generation efficiency and perceptual quality in diffusion-based video systems. Representative works include **IV-Mixed Sampler**, **Golden Noise**, and **CoRe2**.
3. **System translation and industrial deployment.** Beyond academic research, I actively translate algorithmic advances into production-ready systems. Related techniques have been deployed in **Hedra.ai Character-C3**, the **First-Intelligence HYVideo-1.5 API**, and the fundraising demo model for **FuGuang**, giving me end-to-end experience from method design and systems optimization to deployment and product integration.

## Professional Experience

**Oct. 2025 – Present**

**Research Scientist Intern** – First-Intelligence (California Startup).

- **Realtime Video Generation:** Serving as first author and first core contributor on realtime and efficient deployment research for **HYVideo-1.5**, with a focus on **dynamic attention**, **attention sink**, **infinity-DMD**, **infinity-RoPE**, and **self-forcing**.
- **Model Distillation:** Delivered **4-step distillation** for **HYVideo-1.5** and completed **VAE distillation with 4.5× acceleration**, substantially improving efficiency while incurring only limited quality degradation (PSNR from 33 to 32).
- **System Translation:** Contributed these techniques to the production build of the **HYVideo-1.5 API**, supporting user-facing deployment, product delivery, and company fundraising.

**Oct. 2024 – Jul. 2025**

**Research Scientist Intern** – Hedra (California Startup).

- **Talking-Video Foundation Models:** Contributed to **Character-C3**, building a **4-step** digital-human generation model based on **MagicDistillation** for **text+audio+image-to-video** synthesis.
- **Core Distillation Pipeline:** Served as the **sole owner** of distillation deployment, as well as **first author** and **first core contributor**, driving the training, deployment, and productionization of the central distillation pipeline.
- **Industry Translation:** Helped deploy the method stack into the product pipeline and support **Hedra.ai**'s second round of fundraising.

**Jul. 2023 – Mar. 2024**

**Research Intern** – MBZUAI (Zhiqiang Shen's Lab), United Arab Emirates.

- **Dataset Distillation:** Conducted first-author research on dataset distillation, with particular emphasis on more fine-grained **statistical matching** strategies.
- **Representative Projects:** Proposed **G-VBSM** and **EDC**, targeting large-scale data condensation and data optimization, respectively.
- **Research Output: G-VBSM** was selected as a **CVPR Highlight**, while **EDC** was accepted by **NeurIPS**, further consolidating my research line in efficient training and data optimization.

**Oct. 2023 – Dec. 2023**
**Algorithm Intern** – OPPO, Shenzhen, China.

- **Large-Scale Data Condensation:** Designed **G-VBSM** as first author and first core contributor for large-scale data condensation and data optimization.
- **Research Output:** The resulting work was later selected as a **CVPR Highlight** and became one of my representative projects in dataset distillation.

**Dec. 2022 – Oct. 2023**
**AI Systems Intern** – Shanghai AI Laboratory, Shanghai, China.

- **Compiler and Backend Optimization:** Worked on automatic operator optimization based on **TVM** and integrated new backends for high-performance AI systems.
- **Method Exploration:** Proposed **GFlowNet**-based scheduling inference and **Catch-up Distillation**, exploring the intersection of compiler systems and efficient training.

**Sep. 2022 – Jan. 2023**
**Research Intern** – Wu Xinxiao Research Group, Beijing Institute of Technology, Beijing, China.

- **Knowledge Distillation:** Conducted research on knowledge distillation and prepared the resulting paper.
- **Research Output:** The work was accepted as an **IJCAI 2023 Oral**, forming one of my earliest representative projects in knowledge distillation.

**Jul. 2022 – Nov. 2022**
**Algorithm Intern** – Yiliu Technology, Beijing, China.

- **Model Reproduction and Debugging:** Reproduced **Swin-Transformer-V2** and **NeRF** on **OneFlow**, systematically diagnosing missing operators and seed-related precision degradation.
- **Engineering Delivery:** Resolved key implementation issues and accumulated hands-on experience in model engineering and low-level systems debugging.

## Education

Sep. 2024 – Present
- **The Hong Kong University of Science and Technology (Guangzhou)**, Ph.D. Student in Data Science and Analytics.
  Research Focus: efficient AIGC, video diffusion distillation, and sampling optimization; Advisor: Zeke Xie.

Jun. 2024 – Sep. 2024
- **Southeast University**, Professional Master's Degree in Electronic Information.
  Research Focus: pattern recognition, computer vision, knowledge distillation, dataset distillation, and diffusion models.

Sep. 2017 – Jun. 2021
- **East China Normal University**, B.Eng. in Computer Science and Technology.

## Selected Publications

- * denotes equal contribution; ✉ denotes corresponding author. Representative works are selected to reflect my research trajectory in efficient generative modeling, video diffusion, and distillation.

**2026**

1. **Shitong Shao**, Z. Zhou, D. Xie, Y. Fang, T. Ye, L. Bai, B. Han, Z. Xie, "Improved and Accelerated Text-to-Image Generation with Collect, Reflect, and Refine," in **TPAMI**, 2026
   Diffusion Models, Efficient Sampling
2. **Shitong Shao**, Y. Gu, Z. Xie, "FastLightGen: Fast and Light Video Generation with Fewer Steps and Parameters," in **CVPR**, 2026 Video Generation, Step and Size Distillation
3. R. Huang*, **Shitong Shao***, Z. Zhou, P. Zhao, H. Guo, T. Ye, L. Bai, S. Yang, Z. Xie, "Accelerating Diffusion Model Training under Minimal Budgets: A Condensation-Based Perspective," in **CVPR**, 2026
   Dataset Condensation

4. **Shitong Shao**, Z. Zhou, H. Li, Y. Song, W. Zhong, L. Bai, Z. Xie, "Lightning Unified Video Editing via In-Context Sparse Attention," in **ICML under review**, 2026 Video Editing, Sparse Attention
5. H. Li, **Shitong Shao**, et al., "PISA: Piecewise Sparse Attention Is Wiser for Efficient Diffusion Transformers," in **ICML under review**, 2026 Efficient Diffusion Transformers
6. L. Bai, Z. Zhou, **Shitong Shao**, et al., "Optimizing Few-Step Generation with Adaptive Matching Distillation," in **ICML under review**, 2026 Few-Step Generation

**2025**

1. **Shitong Shao**, Z. Zhou, L. Bai, H. Xiong, Z. Xie, "IV-Mixed Sampler: Leveraging Image Diffusion Models for Enhanced Video Synthesis," in **ICLR**, 2025 Video Diffusion, Enhanced Sampling
2. Z. Zhou, **Shitong Shao**, et al., "Golden Noise for Diffusion Models: A Learning Framework," in **ICCV**, 2025 Diffusion Model Quality Optimization, Citations: 81
3. H. Yi*, T. Ye*, **Shitong Shao***, X. Yang*, J. Zhao*, et al., "MagicInfinite: Generating Infinite Talking Videos with Your Words and Voice," in **Technical Report**, 2025 Character-C3, Talking-Video Generation
4. H. Yi, **Shitong Shao**, T. Ye, J. Zhao, Q. Yin, M. Lingelbach, L. Yuan, Y. Tian, E. Xie, D. Zhou, "Magic 1-for-1: Generating One-Minute Video Clips within One Minute," in **Technical Report**, 2025 Long-Video Generation
5. B. Liu, **Shitong Shao**, B. Li, L. Bai, H. Xiong, J. Kwok, S. Helal, Z. Xie, "Alignment of Diffusion Models: Fundamentals, Challenges, and Future," in **arXiv**, 2025 Diffusion Model Alignment Survey
6. **Shitong Shao**, Z. Shen, L. Gong, H. Chen, X. Dai, "Precise Knowledge Transfer via Flow Matching," in **arXiv**, 2025 Knowledge Distillation

**2024**

1. **Shitong Shao**, Z. Yin, M. Zhou, X. Zhang, Z. Shen, "Generalized Large-Scale Data Condensation via Various Backbone and Statistical Matching," in **CVPR Highlight**, 2024 Dataset Condensation
2. **Shitong Shao**, Z. Zhou, H. Chen, Z. Shen, "Elucidating the Design Space of Dataset Condensation," in **NeurIPS**, 2024 Dataset Condensation
3. H. Chen, Y. Dong, **Shitong Shao**, Z. Hao, X. Yang, H. Su, J. Zhu, "Your Diffusion Model is Secretly a Certifiably Robust Classifier," in **NeurIPS**, 2024 Diffusion Model Theory
4. Z. Zhou, Y. Shen, **Shitong Shao**, H. Chen, L. Gong, S. Lin, "Rethinking Centered Kernel Alignment in Knowledge Distillation," in **IJCAI Oral**, 2024 Knowledge Distillation
5. H. Sun, L. Li, P. Dong, Z. Wei, **Shitong Shao**, "Auto-DAS: Automated Proxy Discovery for Training-free Distillation-aware Architecture Search," in **ECCV**, 2024 Distillation-aware Architecture Search

**2023 and Earlier**

1. **Shitong Shao**, H. Chen, Z. Huang, L. Gong, S. Wang, X. Wu, "Teaching What You Should Teach: A Data-Based Distillation Method," in **IJCAI Oral**, 2023 Knowledge Distillation
2. **Shitong Shao**, X. Yuan, Z. Huang, Z. Qiu, S. Wang, K. Zhou, "Expanding Dataset for 2D Medical Image Segmentation using Diffusion Models," in **IJCAI Workshop**, 2023 Diffusion Model Applications
3. M. Zhou, Z. Yin, **Shitong Shao**, Z. Shen, "Self-supervised Dataset Distillation: A Good Compression Is All You Need," in **arXiv**, 2023 Dataset Distillation
4. **Shitong Shao**, X. Dai, S. Yin, L. Li, H. Chen, Y. Hu, "Catch-up Distillation: You Only Need to Train Once for Accelerating Sampling," in **arXiv**, 2023 Diffusion Model Distillation
5. W. Li, **Shitong Shao**, W. Liu, Z. Qiu, Z. Zhu, W. Huan, "What Role Does Data Augmentation Play in Knowledge Distillation?," in **ACCV Oral**, 2022 Knowledge Distillation
6. H. Chen, **Shitong Shao**, Z. Wang, Z. Shang, J. Chen, X. Ji, X. Wu, "Bootstrap Generalization Ability from Loss Landscape Perspective," in **ECCV Workshop**, 2022 Generalization
7. W. Li, **Shitong Shao**, Z. Qiu, A. Song, "Multi-perspective Analysis on Data Augmentation in Knowledge Distillation," in **Neurocomputing**, 2022 Knowledge Distillation
8. J. Xie, L. Gong, **Shitong Shao**, S. Lin, L. Luo, "Hybrid Knowledge Distillation from Intermediate Layers for Efficient Single Image Super-Resolution," in **Neurocomputing**, 2022 Knowledge Distillation

**Full publication list:** Google Scholar profile:
https://scholar.google.com/citations?user=hmUOaNcAAAAJ&hl=en, 56 publications in total.

## Honors and Awards

- Aug. 2022 **Third Place**, ECCV 2022 Workshop Competition on Context-General Domain Generalization; Aug. 2022 **Third Prize**, Jiangsu Provincial Mathematical Contest in Modeling; 2023 **Top 64**, Huawei Software Elite Challenge 2023; May 2022 **Third Prize**, Jiangsu Provincial Data Mining Competition; May 2021 **First Prize**, Computer Skills Application Competition; Apr. 2021 **Meritorious Winner (M Award)**, Mathematical Contest in Modeling; Mar. 2021 **First Prize**, Shanghai Mathematics Competition; Sep. 2020 **First Prize (Shanghai Division)**, Chinese Mathematics Competition.

## Research Summary

I am an AI researcher focusing on video diffusion distillation, sparse attention, and efficient generative modeling. My work examines how to improve video diffusion models across multiple efficiency scales, including **step reduction**, **model compression**, **realtime inference**, and **VAE acceleration**, while also studying how sampling-trajectory design can improve generation quality and lower deployment cost. My recent research and engineering efforts can be summarized along three closely connected directions:

1. **Efficient sampling and efficient generation.** I have systematically worked on step distillation, size distillation, realtime distillation, VAE distillation, and sparse attention, forming a coherent technical line for the efficient deployment of video diffusion models.
   - **Step distillation:** I have continuously advanced few-step video generation through **MagicDistillation**, **FastLightGen**, and **AMD**. Among them, **MagicDistillation** was one of my earliest systematic attempts to adapt **DMD** to video generation and belongs to the early wave of work introducing DMD into video generation models; it is currently under review at **TPAMI**. Building on this line, **AMD** further integrates reinforcement learning and studies a more refined coupling between **reward modeling** and **DMD**; the work has been submitted to **ICML**.
   - **Size distillation:** I proposed a joint step-and-size distillation framework, leading to **FastLightGen**, which was accepted by **CVPR 2026**.
   - **Realtime distillation:** At **First-Intelligence**, I have advanced realtime video generation for **HYVideo-1.5**, systematically exploring dynamic attention, attention sink, infinity-DMD, infinity-RoPE, and self-forcing.
   - **VAE distillation:** Also at **First-Intelligence**, I completed efficient distillation of the **HYVideo-1.5 VAE**, achieving **4.5**× acceleration and substantially reducing deployment cost under acceptable quality loss.
   - **Sparse attention:** I have also worked on sparse attention mechanisms through **PISA** (second author, under review at **ICML**) and **LIVEditor** (first author, under review at **ICML**), focusing on efficient diffusion transformers and sparse contextual modeling for unified video editing.

   These efforts were primarily driven by me as **first author** or **first core contributor**, with the shared objective of improving sampling efficiency, reducing inference cost, and making high-quality video diffusion systems more deployable in practice.
2. **Sampling-trajectory optimization and generation quality.** I also study how initial-noise design and trajectory control can improve generation quality in diffusion-based video synthesis.
   - **IV-Mixed Sampler:** This earlier work improves video generation quality by coupling image diffusion models with video diffusion models, and was accepted by **ICLR 2025**.
   - **Golden Noise:** This project focuses on optimizing sampling initial conditions by learning high-quality initial noise, and was accepted by **ICCV 2025**.
   - **CoRe2:** Going beyond the single-noise setting in **Golden Noise**, **CoRe2** learns optimized noise along the sampling trajectory, providing a more systematic trade-off between efficient sampling and quality-oriented sampling; it was accepted by **TPAMI 2026**.

   These projects were likewise developed with my direct involvement as a **core contributor**.
3. **Industrial deployment and system translation.** I actively push research outcomes into real video-generation systems and product capabilities.
   - **Hedra.ai:** My work **MagicDistillation** was adopted as the distillation solution behind **Character-C3** and deployed successfully in production, contributing to the company's second round of fundraising.
   - **FuGuang (FuLi.ai):** Together with my Ph.D. advisor **Zeke Xie**, I helped support the first round of fundraising through a demo model built on **MagicDistillation** and **FastLightGen**. The system

post-trained a **Wan2.2 low-noise model**, preserved roughly **70%** of the original parameter count, enabled **4-step** video generation, produced **540p** outputs first, and then upsampled them to **720p** with **FlashVSR**, achieving **5-second video generation within 7 seconds**.

- **First-Intelligence:** My **VAE distillation** pipeline has been integrated into the **HYVideo-1.5** generation API. We also trained a **4-step** version of HYVideo-1.5 and combined it with optimization based on the **Muon** optimizer, the **UniPercept reward function**, and a **4.5**×-accelerated VAE to substantially reduce the cost of video generation deployment.
- **Video Agent:** I contributed to related capability building, including a script-analysis skill that already supports script rewriting for user-specified IP.

Across these translational projects, I served as the **first core contributor**.