Ngoc-Son Nguyen

✉ nngocson01042002@gmail.com

Website — nngocson — Google Scholar — nngocson2002

RESEARCH INTERESTS

Multimodal AI, Generative AI, Speech Processing, Audio-Visual Speech Generation, Visual Voice Cloning, Video-to-Speech Synthesis

EDUCATION

University of Science - Vietnam National University HCMC (HCMUS-VNU), Vietnam — Sep 2020 - Sep 2024
Bachelor of Science in Data Science

Graduated with **Highest Distinction** (GPA: 9.27/10.00, 3.86/4.00 – Ranked **2nd** in the Class of 2024)
Thesis: Advancing Vietnamese Visual Question Answering with Transformer and Convolutional Integration (Score: **10.0/10.0**)
Thesis Supervisor: Dr. Tung Le

PUBLICATIONS & PREPRINTS

(*): Equal contribution

[P1] **Ngoc-Son Nguyen**, Thanh V. T. Tran, Hieu-Nghia Huynh-Nguyen, Truong-Son Hy, Van Nguyen. DiFlowDubber: Discrete Flow Matching for Automated Video Dubbing via Cross-Modal Alignment and Synchronization, (*Under Review*)

[P2] **Ngoc-Son Nguyen**, Thanh V. T. Tran, Jeongsoo Choi, Hieu-Nghia Huynh-Nguyen, Truong-Son Hy, Van Nguyen. DiFlow-TTS: Compact and Low-Latency Zero-Shot Text-to-Speech with Factorized Discrete Flow Matching, (*Under Review*)

[P3] Thanh V. T. Tran, **Ngoc-Son Nguyen**, Luong Tran, Long-Khanh Pham, Paarth Neekhara, Shehzeen Samarah Hussain, Van Nguyen. Precise Video-to-Audio Generation with Cross-Modal Alignment in Latent Space, (*Under Review*)

[P4] Hieu-Nghia Huynh-Nguyen, Huynh Nguyen Dang, **Ngoc-Son Nguyen**, Van Nguyen. Flamed-TTS: Flow Matching Attention-Free Models for Efficient Generating and Dynamic Pacing Zero-shot Text-to-Speech, (*Under Review*)

[P5] Cuong Tran Van*, Trong-Thang Pham*, **Ngoc-Son Nguyen**, Duy Minh Ho Nguyen, Ngan Le. DuFal: Dual-Frequency-Aware Learning for High-Fidelity Extremely Sparse-view CBCT Reconstruction, *Transactions on Machine Learning Research* (*TMLR 2025*)

[P6] Trong-Thang Pham, AKASH AWASTHI, Saba Khan, Esteban Duran Marti, Tien-Phat Nguyen, Khoa Vo, Minh Tran, **Ngoc-Son Nguyen**, Cuong Tran Van, Yuki Ikebe, Anh Totti Nguyen, Anh Nguyen, Zhigang Deng, Carol C. Wu, Hien Van Nguyen, Ngan Le. CT-ScanGaze: A Dataset and Baselines for 3D Volumetric Scanpath Modeling , *Proceedings of the IEEE/CVF International Conference on Computer Vision 2025 (ICCV 2025) (**Highlight**)*

[P7] Hieu-Nghia Huynh-Nguyen, **Ngoc-Son Nguyen**, Huynh Nguyen Dang, Thieu Vo, Truong-Son Hy, Van Nguyen. OZSpeech: One-step Zero-shot Speech Synthesis with Learned-Prior-Conditioned Flow Matching , *Proceedings of the 63nd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*

[P8] Khai Le-Duc*, Ryan Zhang*, **Ngoc-Son Nguyen***, Tan-Hanh Pham, Anh Dao, Ba Hung Ngo, Anh Totti Nguyen, Truong-Son Hy. LiteGPT: Large Vision-Language Model for Joint Chest X-ray Localization and Classification Task, (*Preprint*)

[P9] **Ngoc-Son Nguyen**, Van Son Nguyen, Tung Le. Advancing Vietnamese Visual Question Answering with Transformer and Convolutional Integration, *Computers and Electrical Engineering (Q1, IF = 4.9)*

RESEARCH EXPERIENCE

AI Center, FPT Software — Ho Chi Minh, Vietnam
AI Research Resident — Aug 2024 - Present

Supervisor: Dr. Van Nguyen, Prof. Truong-Son Hy, Prof. Ngan Le
Research topics: Multimodal AI, Generative AI, Speech Processing, Medical Imaging

- Proposed a novel vision-language model tailored for medical imaging, specifically for joint localization and classification tasks on chest x-rays. The unified framework integrates multiple pre-trained visual encoders to enhance diagnostic accuracy. Established a strong baseline for future research, conducted ablation studies, and performed comprehensive result analysis. **[P8] (*arXiv 2024*)**
- Proposed an idea of using a learned prior as initialization instead of starting from Gaussian noise, followed by a rectified flow transformation to reach the target distribution, demonstrating improved effectiveness. Implemented the Montreal Forced Aligner to align phonemes with audio, and conducted experiments that yielded valuable insights into emerging challenges. **[P7] (*ACL 2025*)**
- Conducted extensive experiments to demonstrate the advantages of the proposed dataset for 3D volumetric scanpath modeling. Proposed a dual-encoding architecture with a Frequency-Enhanced Dual-Encoding design, which processes X-ray projections through parallel Spatial and Frequency encoders. This approach mitigates the inherent low-frequency bias of conventional CNN-based methods, enabling the model to simultaneously capture global structural context and fine-grained high-frequency anatomical details (e.g., pulmonary nodules and mandibular canals) that are critical for medical imaging. **[P6] (ICCV 2025), [P5] (TMLR 2025)**
- Conducted extensive benchmarking experiments comparing the proposed method with state-of-the-art baselines, including effective prompt design for SoundCap dataset construction, data preprocessing, and comprehensive evaluation to assess performance and

robustness. **[P4]** **(*arXiv 2025*)**, **[P3]** **(*Under Review*)**

- Proposed DiFlow-TTS, a novel zero-shot text-to-speech (TTS) system that investigates discrete flow matching as a viable direction for generative speech modeling. This work serves as an entry point for further research in discrete flow–based speech generation. Extensive empirical evaluations analyze both the strengths and limitations of the approach across key aspects, including naturalness, expressive attributes, speaker identity, and inference latency. **[P2]** **(*Under Review*)**
- Proposed DiFlowDubber, a video dubbing framework with a two-stage training pipeline that effectively transfers knowledge from a pre-trained text-to-speech (TTS) model to video-driven dubbing. The method employs a discrete flow matching generative backbone to address common limitations of TTS-based adaptation pipelines, including weak expressive prosody, limited acoustic richness, and imprecise audio-visual synchronization. **[P1]** **(*Under Review*)**

Multi-modal Machine Comprehension Group, FIT-HCMUS Ho Chi Minh, Vietnam
Research Assistant Sep 2023 - Jun 2024

Research topics: Vietnamese Visual Question Answering

- Proposed the integration of Transformer architectures with Convolutional Neural Networks to improve image information extraction. Studied effective fusion techniques for visual and textual information, and conducted extensive research to optimize model performance on the ViVQA dataset, achieving state-of-the-art results. **[P9]** **(*Journal Computers and Electrical Engineering 2024*)**

HONORS & AWARDS

Outstanding Graduate Award, Vietnam National University, Ho Chi Minh City (VNU-HCMC)	Dec 2024
Outstanding Graduate in Data Science Award, University of Science, VNU-HCMC (HCMUS)	Dec 2024
Outstanding Achievement Award in Science and Technology Research, University of Science, VNU-HCMC (HCMUS)	Nov 2024
University Scholarship for Excellent Academic Achievement, University of Science, VNU-HCMC (HCMUS)	Sep 2020 - Sep 2023

TECHNICAL SKILLS

- **Programming:** C/C++, Python.
- **Framework:** PyTorch, Pytorch Lightning, SQL, Scikit-learn, Pandas, Matplotlib.
- **Tools**: Linux, Git/GitHub, LaTeX.
- **Languages**: Vietnamese (Native), English (740/990 TOEIC Listening and Reading; 310/400 TOEIC Speaking and Writing).

REFERENCES

Dr. Van Nguyen
AI LAB Manager
FPT Software AI Center, Vietnam
E-mail: vannth19@fpt.com
Scholar Profiles: Google Scholar

Prof. Truong-Son Hy
Tenure-Track Assistant Professor
Department of Computer Science & Heersink School of Medicine
The University of Alabama at Birmingham, United States
E-mail: thy@uab.edu
Scholar Profiles: Personal Page — Google Scholar

Prof. Ngan Le
Assistant Professor, Director of Artificial Intelligence and Computer Vision Lab
Department of CSCE
University of Arkansas, United States
E-mail: thile@uark.edu
Scholar Profiles: Google Scholar

Dr. Tung Le
Lecturer
Department of Knowledge Engineering, Faculty of Information Technology
University of Science - Vietnam National University HCMC, Vietnam
E-mail: lttung@fit.hcmus.edu.vn
Scholar Profiles: Google Scholar