Trung M. Bui, PhD

AI / Computer Vision Engineer — Multimodal Foundation Models

Morgan Hill, CA • +1 669-326-2460 • bmtrungvp@gmail.com
linkedin.com/in/trung-m-bui • github.com/mtbui2010 • aistations.org • U.S. Green Card

Summary

Computer Vision and AI Engineer with 7+ years building production deep learning systems and 11+ years total CV experience including PhD research. Specializes in detection, segmentation, 6D pose estimation, and the integration of vision-language models (VLMs) with perception pipelines for multimodal scene understanding. Experience with VLM applications (CLIP, BLIP, LLaVA, Gemini Vision), RAG systems, prompt engineering, and agent frameworks. Open-source contributor to **PyPlanner**, an LLM task-planning library benchmarked in AI2-THOR with seven planning methods (CoT, ReAct, Self-Refine, Hierarchical, etc.), and to **vision-language-action (VLA)** manipulation research reaching 98.75% success on the LIBERO benchmark. PhD with publications in IEEE TIP (Q1) and IEICE Transactions.

Experience

Computer Vision & AI Engineer *March 2019 – Present*

Korea Electronics Technology Institute (KETI), Seongnam, South Korea

- Built a multimodal AI system integrating vision-language models with perception pipelines, enabling natural-language-driven scene understanding and autonomous task execution from visual input.
- Applied VLMs (CLIP, BLIP, LLaVA, Gemini Vision) for zero-shot object recognition, scene captioning, and semantic grounding in cluttered industrial environments.
- Implemented **RAG (Retrieval-Augmented Generation)** pipelines combining foundation models with internal knowledge bases for context-aware reasoning over visual scenes.
- Developed prompt engineering strategies for LLMs and VLMs to elicit structured outputs from complex multimodal queries, integrated into production decision-making loops.
- Designed an agent framework with tool-using foundation models that coordinates perception modules, reasoning, and downstream execution across multi-step workflows.
- Trained and shipped deep learning models for detection, segmentation, and 6D pose estimation on RGB-D data; improved generalization on unseen objects by 20% with vision transformer architectures.
- Built end-to-end training pipelines: dataset curation, augmentation, hyperparameter tuning, evaluation harnesses, and continuous improvement loops for production deployment.
- Optimized inference with TensorRT and ONNX on NVIDIA Jetson, achieving 30+ FPS real-time performance and 50% latency reduction through quantization and operator fusion.
- Led perception architecture for production systems reaching **92% accuracy** in cluttered real-world scenes; processed hundreds of objects per hour across multiple deployment sites.
- Contributed core perception and multimodal components to national R&D projects with multi-million USD funding.

Selected Projects

Multimodal AI System with VLM-Grounded Reasoning *2024 – Present*

- Built an end-to-end multimodal system combining visual perception models with VLMs and LLMs for scene understanding and decision-making.
- Used VLMs (CLIP-style and instruction-tuned models) for zero-shot recognition, captioning, and visual question answering on industrial scenes.
- Implemented RAG over internal knowledge base; combined with prompting strategies to ground foundation model outputs in verifiable context.
- Designed tool-using agents that integrate perception outputs, foundation model reasoning, and execution interfaces.
- Stack: Python, PyTorch, Hugging Face, LLM/VLM APIs, vector databases, agent frameworks.

Vision-Language-Action (VLA) Manipulation Policies *Open Source*

- Evaluated **OpenVLA-OFT** vision-language-action policies for a Franka Panda arm across two physics simulators (**Isaac Lab** and **MuJoCo**) under a unified, perception-first interface.
- Reached **98.75% overall success on the LIBERO benchmark** (80 episodes; 100% on Spatial/Object/Goal, 95% on Long), edging the 97.1% paper baseline using only 2 trials per task.
- Built reproducible infrastructure: three isolated Conda environments, fine-tuning pipeline (24–30h on 4×A6000, 150K steps), and a LeRobot diffusion-policy baseline conversion.
- Stack: PyTorch, OpenVLA-OFT, Isaac Sim/Lab, MuJoCo, LeRobot, HF Transformers. github.com/mtbui2010/robot_sim_vla

PyPlanner — LLM Task Planning Library for Embodied AI *Open Source*

- Designed a pluggable Python library implementing **seven LLM planning methods** (Direct, CoT, Few-Shot, Self-Refine,

ReAct, Hierarchical, LLM Router) behind a unified interface for embodied task planning.
- Built end-to-end evaluation pipeline in **AI2-THOR**: simulator-grounded ground-truth recording, two-layer goal verification (deterministic + LLM judge), and quantitative benchmarking across plan quality, efficiency, and robustness.
- Backend-agnostic core supporting Ollama (local), OpenAI, and Anthropic through a unified `LLMBackend` interface; deployed live Streamlit demo.
- github.com/mtbui2010/pyplanner • Demo: demo-planner.aistations.org

CareRobotAgent — LangGraph Multimodal Agent with Memory *Open Source*

- Built a **LangGraph**-based agent orchestration system integrating voice (Whisper STT + gTTS), LLM intent routing, task planning, and robot execution in AI2-THOR with auto-replan on failure.
- Implemented dual-memory architecture: **SQLite (episodic) + ChromaDB (semantic)** for personalization and context retrieval based on interaction history.
- Backend-agnostic LLM support (Ollama / OpenAI / Anthropic) configurable via env vars; Streamlit chat UI for interactive demos.
- github.com/mtbui2010/carerobotagent

Fine-Aware Vision Transformer for Precision Detection *Published, IEICE 2025*

- Designed a novel transformer architecture for precision object detection and pose estimation on cluttered RGB-D scenes.
- Achieved **20% precision improvement** over CNN baselines; trained on large-scale dataset with custom augmentation strategy.
- Deployed on NVIDIA Jetson with TensorRT, sustaining 30+ FPS at production accuracy.

Real-Time RGB-D Perception Pipeline *2020 – 2024*

- End-to-end perception stack: 2D detection → segmentation → 6D pose estimation → downstream planning.
- Optimized for edge deployment using TensorRT and ONNX; sustained 30+ FPS on NVIDIA Jetson AGX.
- Reduced perception failure rate by 35% through joint optimization across stages and learned scoring.

Tracker Lab — Multi-Object Tracking from Scratch *Open Source*

- Implemented four MOT algorithms (**SORT, DeepSORT, ByteTrack**, and a custom variant) from scratch — Kalman filter, Hungarian assignment, and ReID matching cascade — with unit tests.
- Detection via fine-tuned **YOLOv11**; appearance embeddings via **OSNet** ReID; custom **MOTA / IDF1 / HOTA** metrics benchmarked against MOT17/MOT20/DanceTrack.
- Full-stack platform: FastAPI orchestration, Next.js 14 + Canvas UI, WebSocket GPU workers, Docker / Vercel / RunPod serverless deployment. github.com/mtbui2010/vision_tracking

Single Image Dehazing Using Color Ellipsoid Prior *IEEE TIP 2018*

- Developed a novel algorithm for single-image dehazing as part of PhD research.
- Published in IEEE Transactions on Image Processing, a top-tier journal in computer vision.
- Outperformed existing methods on standard benchmarks across diverse atmospheric conditions.

Selected Publications

T. M. Bui, J. Hwang, S. Jun, W. Kim, D. Shin. "A Fine-Aware Vision Transformer for Precision Grasp Pose Detection." *IEICE Transactions on Information and Systems*, Nov. 2025.

T. M. Bui, Y. Kim, S. J. Moon, M. Cho, M. Seo, D. Shin. "Development of a Mobile Assistive Robot for Daily Living Support." *Ubiquitous Robots*, 2025.

T. M. Bui, W. Kim. "Single Image Dehazing Using Color Ellipsoid Prior." *IEEE Transactions on Image Processing*, Feb. 2018. (Q1 journal)

T. M. Bui, H. N. Tran, W. Kim, S. Kim. "Segmenting Dark Channel Prior in Single Image Dehazing." *IET Electronics Letters*, March 2014.

Education

Ph.D. in Computer Vision, Kyung Hee University, South Korea 2014 – 2019
Thesis: *Single Image Dehazing Using Color Ellipsoid Prior* (basis of IEEE TIP 2018 publication)

M.Eng. in Computer Vision, Kyung Hee University, South Korea 2011 – 2014

B.Eng. in Electrical & Electronics Engineering, Ho Chi Minh City University of Technology, Vietnam 2005 – 2010
Excellent Engineer Training Program

Technical Skills

Multimodal AI & Foundation Models	VLMs (CLIP, BLIP, LLaVA, Gemini Vision, OWL-ViT), LLM integration, prompt engineering, RAG, agent frameworks, multi-step reasoning, multimodal grounding
LLM Orchestration	LangGraph, LangChain, tool-using agents, LLM task planning (CoT, ReAct, Self-Refine, Hierarchical)
Computer Vision	object detection, semantic segmentation, 6D pose estimation, scene understanding, RGB-D perception, video understanding
Deep Learning	CNN architectures, vision transformers, attention mechanisms, transfer learning, self-supervised learning, hyperparameter tuning
ML Frameworks	PyTorch, TensorFlow, Hugging Face, ONNX, TensorRT, CUDA
Programming	Python (expert), C++, C
Data & ML Engineering	large-scale training pipelines, dataset curation, augmentation, evaluation harnesses, continuous improvement loops, model benchmarking
Memory & Retrieval	ChromaDB, SQLite, vector databases, semantic search, RAG patterns
Deployment	TensorRT optimization, ONNX, NVIDIA Jetson, real-time inference (30+ FPS), model quantization & pruning
Infrastructure	Docker, Linux (Ubuntu), Git, distributed training

Certifications

Visual Perception for Self-Driving Cars — Coursera (2023) • Improving Deep Neural Networks — DeepLearning.AI (2021)