Hongyang Du

61 Pitman St, Providence, RI | 240-413-5747 | hongyangdu182@gmail.com | Website | LinkedIn | Github

Image/Video Diffusion Models | World Models | Representation Learning

Education

Brown University **Aug 2025 – May 2027**
M.S. Computer Science (AI/ML Track) GPA 4.0/4.0

University of Maryland, College Park **Aug 2021 – May 2025**
B.S. Computer Science (Honors), Minor in Mathematics GPA 4.0/4.0

Professional Experience

Adobe-Machine Learning Engineer Intern | Content AI Team May 2026 – Aug 2026

- Built on production, consumer-facing agentic image creation: a frozen LLM composes designs by driving commercial Adobe software across 230+ tools and up to 80 tool-calling iterations.
- Designed a self-improving system that treats the skill bank as the policy: distilled user trajectories and synthesized briefs into natural-language workflows to make agent more capable, more reliable, and faster.
- Achieved +17.6% pairwise win rate over the baseline across four graphic-design benchmarks and lifted success rate from 72.7% to 99.3% on GenEval2, at 3.4–6.2% latency overhead.

iFlytek-Machine Learning Engineer Intern | Agriculture MLLM Team June 2024 – Aug 2024

- Pretrained a 2B Chinese LLM from scratch based on the LLaMA2 architecture (e.g., RoPE positional encoding, RMSNorm, SwiGLU). Employed SentencePiece Unigram tokenizer to enhance Chinese linguistic modeling.
- Combined ResNet hierarchical classification and YOLOv10 disease-spot detection to classify over 1,000 crop diseases and pests. Constructed RAG-enhanced prompts using visual clues and a knowledge graph.
- Finetuned the model using LoRA and reinforcement learning for domain-specific adaptation. Build a Chinese-centric pretraining pipeline with support for FSDP, gradient checkpointing, and customized preprocessing for Chinese text.

Visual Representation & Generation

Unified Representation Autoencoder for Image Diffusion **Present**
Under Review

- Conventional Representation Autoencoders (RAEs) exhibit a structural discrepancy, as reconstruction overfits to pixel features while diffusion overfits to deep semantic features, leaving cross-frequency mutual information isolated.
- We propose a self-supervised training paradigm that replaces heuristic feature-fusion, mathematically proven to capture joint correlations between pixel-level details and semantic features.
- By harmonizing reconstruction and generation in an optimized latent space, a unified decoder improves reconstruction fidelity and accelerates DiT convergence by 2x with lower gFID.

Video World Model **Feb 2026**
ICML 2026, CVPR VGBE Challenge [3rd Place] & Workshop [Oral]

- Video Diffusion Models inherently struggle with temporal inconsistency and spatial drift, largely due to a lack of robust grounding in 3D structural priors and physical geometric laws.
- We proposed a self-supervised data framework to distill 3D priors from a Geometry Foundation Model (VGGT). Leveraging Direct Preference Optimization (DPO) for physical geometric laws alignment.
- With minimal preference data and RL fine-tuning, our framework significantly outperforms state-of-the-art baselines in 3D consistency and motion coherence, achieving superior physical plausibility on complex real-world benchmarks.

Implicit Neural Representation for Video Compression **May 2025**
WACV 2026

- Conducted empirical study of video compression methods based on INRs. Developed a unified benchmarking platform to evaluate models across metrics PSNR, MS-SSIM, bitrate, training time, and decoding speed on HD datasets.
- Proposed a novel architecture RNeRV, which introduced a Weight Token Masking mechanism and HyperNetwork decoupling to achieve flexible bitrate control and efficient model deployment. Demonstrated +1.27% average improvement in visual quality under equal training budgets and enhanced real-time inference throughput.
- Built the XINC (eXplainable INR Compression) tool to visualize neuron-level activation and contribution maps across layers and time steps, improving interpretability and facilitating model optimization for video streaming.

VLMs & Agents

Self-Evolving Vision Language Models From Zero Data **June 2026**
EMNLP 2026

- Proposed a novel training paradigm that eliminates the need for human-annotated data ("zero data") by designing an innovative three-role collaborative pipeline: Proposer, Coder, and Solver
- Empowered the model to autonomously construct visually complex scenes via code generation; implemented Group Relative Policy Optimization (GRPO) with role-specific reward mechanisms for sequential, closed-loop training.
- Drive a comprehensive performance leap across base models (e.g., Qwen3-VL, Mimo-VL) with significant gains in multimodal understanding, mathematical reasoning, and coding capabilities.

Agent Skill Dependency-Aware Structural Retrieval **Apr 2026**
EMNLP 2026

- Structured massive agent skill libraries as an offline dependency graph and retrieved small, relevant skill bundles at inference (seeding + PageRank under a token budget) instead of loading the full library.
- Reduced context tokens while improving downstream reward for LLM-based agents operating over large skill sets.

Vision Language Models Understanding and Hallucination **Aug 2025**
NeurIPS 2025, **CVPR 2025 Workshop [Oral]**

- Designed VideoHallu dataset for hallucination detection in synthetic video understanding; proposed 2 novel metrics (temporal consistency, semantic mismatch) to quantify alignment between captions and dynamic video scenes.
- Implemented robust hallucination detection framework integrating frame-level caption matching, embedding similarity (CLIP/BLIP), and cross-frame alignment; enhanced via GRPO post-training with curriculum learning on video QA datasets incorporating counterintuitive commonsense and physics reasoning.
- Conducted a systematic survey of cutting-edge vision-language models (e.g., GPT-4V, Gemini, Qwen-VL), detailing their model architectures, instruction alignment mechanisms, and evaluation protocols.

Publications

MM-Zero: Self-Evolving Multi-Model Vision Language Models From Zero Data
Z. Li*, **H. Du***, C. Huang*, et al. · **EMNLP 2026**

VideoGPA: Distilling Geometry Priors for 3D-Consistent Video Generation
H. Du*, J. Ye*, X. Cong*, et al. · **ICML 2026**

Graph of Skills: Dependency-Aware Structural Retrieval for Massive Agent Skills
D. Liu*, Z. Li*, **H. Du**, et al. · **EMNLP 2026**

A Cookbook of 3D Vision: Data, Learning Paradigms, and Application
H. Du*, Z. Li*, D. Liu*, et al. · **CVPR 2026 Workshop**

How to Design and Train Your Implicit Neural Representation for Video Compression
M. Gwilliam, R. Zhang, N. Padmanabhan, **H. Du**, et al. · **WACV 2026**

VideoHallu: Evaluating and Mitigating Multi-modal Hallucinations for Synthetic Videos
Z. Li*, X. Wu*, Y. Qin, **H. Du**, et al. · **NeurIPS 2025**

A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges
Z. Li*, X. Wu*, **H. Du**, et al. · **CVPR 2025 Workshop (Oral)**

Ipelets for the Convex Polygonal Geometry
N. Parepally, A. Chatterjee, A. Gezalyan, **H. Du**, et al. · **SoCG 2024**

Technical Skills & Award

Languages: Python, Shell, Java, C++, Bash, Go, Rust, SQL, JS/HTML/CSS, OCaml, Lua, Racket, React, Matlab
Dev Stack: Spring Boot, Flask, Django, FastAPI, Node.js, RESTful APIs, Redis, MongoDB, Docker, Flutter, Next.js
AI/ML Stack: PyTorch, TensorFlow, Transformers, scikit-learn, OpenCV, LangChain, LlamaIndex, Haystack, NumPy
Awards: Robert Ma Scholarship, Break Through Tech Scholarship, Dean's List (7 semesters), AIME Qualifier (Top 5%), AMC 12 Top 1% Honor Roll of Distinction, TN State Math Competition 1st Place