Research Vision and Future Directions

Chi Keung Tang

Research Vision: From Multimodal Foundations to Spatiotemporal Agentic AI

Our research program operates at the vanguard of **Spatiotemporal Spatial Intelligence, Multimodal Foundation Models, and Photorealistic 3D/4D World Generation**. Over our career, we have dedicated ourselves to transforming how intelligent systems perceive, interact with, and synthesize the 3D and 4D physical world, bridging the traditional boundary between structural geometry and cognitive reasoning.

Our group focuses on the next major paradigm shift in computer science: transitioning from static 2D digital vision to interactive, physics-grounded 4D spatial intelligence. By orchestrating interactions between Large Language Models (LLMs), vision foundation models, and spatial-temporal representations, our laboratory contributes training-free adaptive protocols, parameter-efficient adapters, and direct dense mapping formulations to solve complex, open-world problems in autonomous robotics, cinematic creation, and physical human-computer simulation.

1. ChatCam, Trace Anything & Point4Cast for 4D World Generation

Traditional 3D environment synthesis imposes immense labor overhead and relies on manually intensive expert pipelines. To democratize this domain, we introduced *WorldCraft*, an agentic framework utilizing LLMs via procedural generation to build complex, photo-realistic indoor and outdoor environments. Within this system, we proposed *ForgeIt* to dynamically master adjustable category parameters via a critic-driven auto-verification manual, alongside *ArrangeIt* to optimize structural layouts under complex ergonomic and functional constraints.

Building on this, our work on *ChatCam* empowers direct camera control through conversational AI. We introduce *CineGPT*, a GPT-based autoregressive model that integrates language understanding with text-conditioned camera trajectory generation, and an *Anchor Determinator* to locate spatial scene objects to anchor exact trajectory positions. To unify semantic intent with multi-view image sets, we developed *Spatially Contextualized VLMs*, injecting a continually updatable working memory composed of a scene portrait, a semantically labeled point cloud, and a high-order scene hypergraph. This framework enables VLMs to construct atmospheric architectural and terrain code in Blender with strict constraint validation.

Complementing trajectory generation, our framework targets the core challenges of continuous spacetime modeling and streaming 4D representations. Through *Trace Anything*, we model continuous spacetime in a single forward pass, defining a dense *Trajectory Field* that assigns every pixel to a spline-based 3D curve, opening up new horizons for fluid 4D scene interpolation. Furthermore, to handle complex visual dynamics dynamically evolving across time, our streaming spatiotemporal architecture, *Point4Cast*, maps sparse point-cloud inputs directly onto structural look-ahead layout forecasts, establishing a solid baseline for scalable lifelong streaming world generation.

In summary, the major contributions in 4D world generation include:

- We introduce *ChatCam*, a system that, for the first time, enables users to operate cameras through natural language interactions, reducing technical hurdles for creators.
- We develop *CineGPT* for text-conditioned camera trajectory generation and an *Anchor Determinator* for precise camera trajectory placement.
- We establish *Trace Anything* and *Point4Cast* for dense trajectory field learning and robust streaming geometry forecasting in complex dynamic maps.

2. Motion-Agent & Zero-Shot Spatiotemporal Reasoning Agents

Teaching machines to interpret and produce dynamic human actions is crucial for modern humanoid platforms. We developed *Motion-Agent*, an efficient conversational framework driving human motion generation,

editing, and bidirectionally aligned captioning. By embedding a trained motion tokenizer/detokenizer with low-rank adapters (LoRA), our generative agent *MotionLLM* maps temporal kinematics to discrete tokens matching an LLM's vocabulary. Coordinated by a central planner, it achieves infinite, long-sequence multi-turn composition and smooth semantic transitions without extensive pre-training.

Furthermore, we have pushed the limits of zero-shot path navigation for multiple coexisting actors in dynamic, cluttered maps. By representing floorplans, obstacles, and agent paths as unstructured text tokens, we demonstrated that advanced reasoning LLMs can act as global path navigators. This architecture calculates routes via sparse anchor points connected by continuous straight lines, and evaluates closed-loop communication update mechanisms (*additive* versus *compositional*) to autonomously resolve real-time collision deadlocks.

To summarize, our contributions include:

- We introduce a simple, efficient conversational framework, *Motion-Agent*, that utilizes pre-trained LLMs and produces state-of-the-art results in various motion-language tasks.
- We demonstrate the flexibility and versatility of our method by achieving highly customizable motion-language tasks, including long and complex motion generation, multi-turn editing, and multi-turn reasoning.

3. Audio-Agent & Physics-Grounded Interaction Fields

To capture the subtle relations that bind dynamic actors together with their surrounding environments, we developed a series of multi-turn visual and sound generation systems:

- **Audio-Agent:** An AI framework that decomposes complex text inputs into step-by-step atomic audio components. It integrates visual and audio streams by fine-tuning an open-source model (*Gemma2-2B-it*) to transform visual inputs into time-aligned semantic tokens, steering a text-to-audio diffusion backbone across long-condition multi-event generation.
- **CoT-Seg & CoT-RVS:** A zero-shot framework extending Chain-of-Thought (CoT) compute to image and video reasoning segmentation. By synthesizing self-correcting meta-queries to rectify spatial mask false positives and computing temporal-semantic keyframe selection profiles in moving video streams, it establishes robust, fine-grained object grounding.
- **SK-Adapter:** A parameter-efficient structural control architecture that treats a 3D skeleton as a first-class spatial token control signal. Injected into a frozen structure generation transformer via cross-attention layers, it accomplishes tight region-specific native 3D editing and articulation while preserving base foundation priors.
- **HA-HOI:** Addressing the physical gap in casual monocular videos, *HA-HOI* implements a *human-first, object-follow* 4D tracking routine. It optimizes semantic contact alignment via VLM contact surface proposals and projects the output into an IsaacGym physics simulator via reinforcement learning to eliminate gravity slipping.

Future Research Agenda

Our laboratory focuses on translating abstract artificial intelligence into embodied spatial domains, leading three critical future thrusts:

1. Physics-Driven Multimodal Simulation Loops. Current vision foundation models are proficient at rendering visual appearance but lack a foundational understanding of intuitive physics. We plan to construct generative frameworks where trajectory fields, volumetric patches, and multi-agent systems are explicitly regularized inside differentiable, closed-loop simulation environments. This ensures that generated virtual worlds are structurally stable, dynamically sound, and safe for synthetic agent training.

2. Lifelong Streaming Spacetime Representations. Expanding on our streaming geometry paradigms (*Point4Cast*), our group will pioneer lifelong, casual world models capable of processing real-time, asynchronous sensory feeds from mobile edge platforms. These systems will maintain a persistently evolving internal latent workspace that handles simultaneous tracking, semantic segmentation, and long-range look-ahead layout forecasting within unpredictable human settings.

3. Sim-to-Real Deployment on Physical Humanoid Systems. We intend to fully capitalize on our human-anchored interaction pipelines to distill scalable behavioral demonstrations out of unstructured, open-world internet data. By forging strong collaborative ties with international hardware and robotics research initiatives, we will compile and deploy these physics-ready interaction vectors onto real humanoid platforms, accelerating robotic manipulation capabilities.