Updated: March 9, 2025

Research Statement

# Toward AI Agentic Modeling and Reasoning: VLM, VLR, SLR, ALM

## 1 Visual-Language Modeling (VLM)

### 1.1 Multimodal Generation of Animatable 3D Human Models with AvatarForge

We introduce *AvatarForge*, a framework for generating animatable 3D human avatars from text or image inputs using AI-driven procedural generation. While diffusion-based methods have made strides in general 3D object generation, they struggle with high-quality, customizable human avatars due to the complexity and diversity of human body shapes, poses, exacerbated by the scarcity of high-quality data. Additionally, animating these avatars remains a significant challenge for existing methods. AvatarForge overcomes these limitations by combining LLM-based commonsense reasoning with off-the-shelf 3D human generators, enabling fine-grained control over body and facial details. Unlike diffusion models which often rely on pre-trained datasets lacking precise control over individual human features, AvatarForge offers a more flexible approach, bringing humans into the iterative design and modeling loop, with its auto-verification system allowing for continuous refinement of the generated avatars, and thus promoting high accuracy and customization. Our evaluations show that AvatarForge outperforms state-of-the-art methods in both text- and image-to-avatar generation, making it a versatile tool for artistic creation and animation.

### 1.2 WorldCraft: 3D World Creation and Customization via LLM Agents

Creating virtual worlds has applications across various domains, but traditional 3D modeling often requires the expertise of trained professionals. To make this process more accessible, we introduce *WorldCraft*, a system that leverages large language model (LLM) agents and procedural generation to create both indoor and outdoor scenes populated with customizable objects. Users can manipulate object attributes and scene layouts through intuitive natural language commands. In our framework, a coordinator agent manages the overall process and works with two specialized LLM agents to complete the scene creation: *ForgeIt*, which integrates an ever-growing manual through auto-verification to enable precise customization of individual objects, and *ArrangeIt*, which formulates hierarchical optimization problems to achieve a layout that balances ergonomic and aesthetic considerations. Additionally, our pipeline incorporates a trajectory control agent, allowing users to animate the scene and operate the camera via natural language. Our system is also compatible with off-the-shelf deep 3D generators to enrich scene assets. Through evaluations and comparisons with state-of-the-art methods, we demonstrate the versatility of WorldCraft, ranging from single-object customization to intricate, large-scale interior and exterior scene designs. This system empowers non-professionals to bring their creative ideas to life.

## 2 Visual-Language Reasoning (VLR)

### 2.1 ThinkFirst: High-Quality Reasoning Segmentation with Chain of Thoughts

Reasoning segmentation is a challenging vision-language task that aims to output the segmentation mask with respect to a complex, implicit, and even non-visual query text. Previous works incorporated multimodal Large Language Models (MLLMs) with segmentation models to approach the difficult problem. However, their segmentation quality often falls short in complex cases, particularly when dealing with out-of-domain objects with intricate structures, blurry boundaries, occlusions, or high similarity with surroundings. In this paper, we introduce *ThinkFirst*, a training-free reasoning segmentation framework that leverages GPT's chain of thought to address these challenging cases. Our approach allows GPT-4o or other powerful MLLMs to

generate a detailed, chain-of-thought description of an image. This summarized description is then passed to a language-instructed segmentation assistant to aid the segmentation process. Our framework allows users to easily interact with the segmentation agent using multimodal inputs, such as easy text and image scribbles, for successive refinement or communication. We evaluate the performance of ThinkFirst on diverse objects. Extensive experiments show that, this zero-shot-CoT approach significantly improves the vanilla reasoning segmentation agent, both qualitatively and quantitatively, while being less sensitive or critical to user-supplied prompts after Thinking First.

# 3 Spatial-Language Reasoning (SLR)

## 3.1 Dynamic Path Navigation for Motion Agents with LLM Reasoning

Large Language Models (LLMs) have demonstrated strong generalizable reasoning and planning capabilities. However, their efficacies in spatial path planning and obstacle-free trajectory generation remain underexplored. Leveraging LLMs for navigation holds significant potential, given LLMs' ability to handle unseen scenarios, support user-agent interactions, and provide global control across complex systems, making them well-suited for agentic planning and humanoid motion generation. As one of the first studies in this domain, we explore the zero-shot navigation and path generation capabilities of LLMs by constructing a dataset and proposing an evaluation protocol. Specifically, we represent paths using anchor points connected by straight lines, enabling movement in various directions. This approach offers greater flexibility and practicality compared to previous methods while remaining simple and intuitive for LLMs. We demonstrate that, when tasks are well-structured in this manner, modern LLMs exhibit substantial planning proficiency in avoiding obstacles while autonomously refining navigation with the generated motion to reach the target. Further, this spatial reasoning ability of a single LLM motion agent interacting in a static environment can be seamlessly generalized in multi-motion agents coordination in dynamic environments. Unlike traditional approaches that rely on single-step planning or local policies, our training-free LLM-based method enables global, dynamic, closed-loop planning, and autonomously resolving collision issues.

# 4 Audio-Language Modeling (ALM)

## 4.1 ReelWave: A Multi-Agent Framework Toward Professional Movie Sound Generation

Film production is an important application of generative audio, where richer context is provided through multiple scenes. In ReelWave, we propose a multi-agent framework for audio generation inspired by the professional movie production process. We first capture semantically and temporally synchronized 'on-screen' sound by training a prediction model that forecasts three interpretable, time-varying audio control signals: loudness, pitch, and timbre. These three parameters are then specified as conditions by a cross-attention module. Our framework subsequently infers 'off-screen' sound to complement the generation through cooperative interaction between communicative agents. Each agent takes on specific roles similar to those of a movie production team and is supervised by an agent called the Sound Director. Furthermore, we explore the case when the conditional video consists of multiple scenes, a scenario commonly seen in videos extracted from movies of considerable length. As a result, our framework can capture a richer context for audio generation conditioned on video clips extracted from movies.