# Research Proposal

Tong Zhang

**The central goal of my research is to develop intelligent machines capable of performing human-related visual understanding and generation tasks.** In particular, I study two important questions:

- How to interpret multimodal data, thereby bridging the gap between human cognitive processes and machine learning capabilities?
- How to design intelligent machines that can efficiently recreate and edit human-related visual contents?

More specifically, I focus on learning-based algorithm to address the above challenges. In the following, I will highlight my research experience in these topics.

## 1 Interpretability Of Multimodal Learning

**Human-Readable SVG Generation** The primary goal of my work [1] has been to transform complex image data into Scalable Vector Graphics (SVGs) that are interpretable both to advanced AI models and human users. This task involved overcoming the inherent challenges of traditional pixel-based image representations, such as JPEG or PNG, which suffer from scalability issues and limited search engine indexability.

By employing vision language models, we developed a methodology to transform images into SVG format, which preserves the relational properties and context of the original scene, leading to SVGs that are simpler and more interpretable. The success of this method is evident in its ability to facilitate a more logical and accurate depiction of graphical elements, significantly improving the reasoning and interpretability capabilities of SVGs through Large Language Models such as GPT4.

**Narrations and Reasoning for autonomous vehicular Action** In the realm of autonomous vehicular control, our research [3] aimed at providing narrations and reasoning that describe each decision-making step in the control and action of autonomous vehicles. This work was critical in addressing the growing need for transparency and interpretability in automated decision-making. Our efforts resulted in the development of a system that outperforms existing state-of-the-art video captioning models. This research is a substantial step towards making AI systems more transparent and trustworthy, ensuring that these technologies can be deployed safely and effectively in real-world scenarios.

## 2 Recreating And Editing Human-Related Visual Contents

The realm of human-related visual understanding extends beyond mere recognition of human elements within visual environments; it encompasses the recreation of photorealistic content, such as faces and bodies, adding a dynamic layer to a myriad of applications. My research primarily focuses on conditional visual editing, a domain that involves manipulating human-related images or videos. This manipulation is guided by various conditions, including 2D keypoints, 3D mesh models, or textual descriptions. A critical aspect of my work is to enhance the efficiency of the generation process.

An example of this is illustrated in the paper [2] for WACV 2024. Here, we introduced a novel method to create controllable head avatars from a single reference image. This method utilizes point-based neural rendering combined with transformer , facilitating the efficient rendering of human heads with varied expressions, head poses, and camera views. Unlike previous techniques, which often depend on multi-view inputs or videos of the subject for texture reconstruction, our approach significantly simplifies the process. This innovation not only streamlines content creation but also paves the way for real-time animation applications.

## 3 Future Directions

As I look towards the future of my research in human-related visual understanding and reasoning, several key areas stand out as particularly promising:

**Enhancing Multimodal Interpretability** Building on the foundations laid by my work in SVG generation, a significant direction for future research is the enhancement of multimodal interpretability. I intend to explore how large scale Vision-Language foundation model can be fine-tuned to generate even more sophisticated, yet comprehensible, visual representations, particularly in contexts where nuanced understanding is crucial, such as in sensitive medical imaging scenarios.

**Advancements in Conditional Visual Editing** The field of conditional visual editing holds vast potential, especially in the creation and manipulation of dynamic human-related content. As an example, one of my ongoing works on Grounded Entity Replacement is a testament to the innovative strides being made in the field. This project involves the advanced technique of grounded language-image retrieval coupled with text-to-image generation. The objective is to produce semantically coherent image-text pairs, revolutionizing the way visual data is understood and manipulated. This approach not only enhances the realism and applicability of generated images but also ensures the relationship of entities in both text and image are consistent, thereby advancing the field towards more finegrained visual content editing tools.

## References


[1] **Tong Zhang**, Haoyang Liu, Peiyan Zhang, Yuxuan Cheng, and Haohan Wang. Beyond Pixels: Exploring Human-Readable SVG Generation for Simple Images with Vision Language Models. In ***Under Review*** *at The European Conference on Computer Vision (ECCV)*, 2024. Arxiv link

[2] Haoyu Ma, **Tong Zhang**, Shanlin Sun, Xiangyi Yan, Kun Han, and Xiaohui Xie. CVTHead: One-shot Controllable Head Avatar with Vertex-feature Transformer. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. Arxiv link

[3] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, **Tong Zhang**, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. ADAPT: Action-aware Driving Caption Transformer. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. Arxiv link