Abstract

In recent years, generative models, especially diffusion models, have made significant advancements, pushing the boundaries of traditional image synthesis and addressing various application-specific demands in fields such as virtual reality, film production, and fashion design. These technologies enable the automated generation of characters with precise specifications, dramatically improving workflow efficiency and reducing production costs. One of the most transformative advancements in generative models is multimodal generation, which leverages diverse inputs like text prompts, image references, and spatial guidance (e.g., sketches or segmentation maps). This approach opens up broader creative possibilities, enabling flexible design workflows and improving model robustness and adaptability across tasks. As technology evolves, multimodal inputs are poised to revolutionize character design, enhancing both efficiency and creativity in various industries. Building on multimodal generation, this dissertation investigates character creation and design through generative models incorporating diverse inputs like text, stroke data, and structural maps. The core challenge lies in balancing manipulability, convenience, and precision—three essential yet often competing elements in the design process:

• Manipulability refers to the level of control users have over the generative outputs, allowing flexibility in shaping design details.

• Convenience emphasizes the efficiency of the design workflow, stream lining the process to reduce repetitive tasks and focus on higher-level decisions.

• Precision indicates the model's ability to produce outputs that accurately align with user inputs, such as text descriptions or sketches.

Achieving an optimal balance among these factors is complex, as enhancing one often limits the others. This study addresses this trilemma by identifying the primary contradiction in each design scenario and resolving it according to user needs, guided by Marx's theory of contradictions. This approach enables tailored solutions across three design applications, each focusing on different primary tensions within the trilemma:

• Drawing Multi-Age Facial Features for Anime Characters: This task primarily balances precision and manipulability. Designing lively, age-specific facial features in anime characters demands both detailed control and real-time interactivity. I developed an interactive painting assistance system that leverages user-inputted strokes to create facial features with age-specific characteristics. This system ensures continuous interaction between the user's design intent and the generative model, effectively balancing user creativity and model-driven generation.

• Character Pose Design: This task requires balancing precision and convenience. Quickly generating character images that align with specific descriptions is crucial in fields such as advertising and poster
design. Building upon conditional diffusion models like ControlNet, I developed an enhanced end-to-end text-to-image (T2I) generation framework that enables efficient, accurate pose creation. Users can generate custom character poses by combining text inputs with spatial conditions, such as skeletons, facial landmarks, and sketches, to produce diverse and precise poses.

• Character Head Motion Design: Here, the primary focus is balancing convenience and manipulability for video-based motion design. Traditional frame-by-frame animation is often labor-intensive, especially for designing head movements in character animation. I introduced a head motion prediction model, integrated into an imageto-video (I2V) generative framework, to streamline the workflow. This model uses multimodal inputs—trajectory strokes, audio, and reference images—to predict head movements, allowing for intuitive user control over the motion trajectory while reducing the manual effort required for animation.

In conclusion, this dissertation uniquely contributes to advancing multimodal generative models in character design, optimizing user-customizable workflows across varied scenarios. By carefully balancing manipulability, convenience, and precision, this research enhances the applicability of generative models in creative domains, fostering both efficiency and creativity.

Keywords: Character Design, Generative Models, Multimodal Generation, Computer Vision, Human-Computer Interaction.