Personal Statement

Tingbo Hou

My research interests and expertise are in image and video generation. I'm honored for contributing to Google's Imagen 3 and Meta's Movie Gen. As generative models being largely scaled, training and inference become very expensive. This poses challenges in model development for experimenting alternative solutions and downstream applications. Making efficient representation and designs can reduce costs for development and deployment, while paving paths for incubating new ideas that will eventually lead to step changes of GenAI.

Previous Work

Beyond conventional topics of model compression and knowledge distillation, efficient GenAI has a broad coverage, also including one-step diffusion, linear-complexity models, mixture of experts (MoEs), auto encoder, and more. My previous work touches most of these areas.

One-Step Diffusion

Diffusion models are trained with a large number of steps, e.g. 1000. With advanced samplers, the inference can be reduced to about 10 steps. Our UFOGen paper is the *first* work to introduce one-step generation of diffusion models. Instead of Gaussian, it uses GAN to model the distribution of a denoising step. Both the generator and the discriminator are initialized from a pre-trained diffusion model, making the training of UFOGen fast and stable. Later, we proposed EM Distillation that minimizes an approximation of the mode-covering divergence between a pre-trained diffusion teacher model and a latent-variable student model.

MobileDiffusion

MobileDiffusion is a complex of model pruning, one-step distillation, and efficient training. Remarkably, it can generate a 512x512 image on a mobile device in 0.2 seconds. In addition, it designs an efficient training strategy for network search, with a total cost of 512 TPUs for 15 days.

Linear-Complexity Video Generation

LinGen is a linear-complexity model for text-to-video generation. It adopts bidirectional Mamba as the main computation module. To alleviate the long-range decay of Mamba, we introduced rotary-major scan and review tokens. We also add Swin attention in the temporal domain to enhance temporal consistency. LinGen has 4B parameters, and is the *first* model that can directly generate a minute-long video at 512p resolution.

Real-Time Infinite Video Generation

In this work, we propose StreamDiT, which can generate consistent video streaming from text prompts. It is based on a 4B DiT with window attention. At every diffusion step, it outputs a few latent frames and enqueues the same number of new latent frames for denoising. The generation is consistent, taking prompts on-the-fly. StreamDiT can generate streaming videos at 16 fps on one GPU.

Transformer Auto-Encoder

Conventional auto-encoders for images and videos are based on CNNs, which compress visual content uniformly in space and temporal domains. Yet, images and videos are highly non-uniform. In the work of ViToK, we studied the design of using transformers for auto-encoding. It can achieve comparable quality with CNN based auto-encoders, while being more efficient and scalable in computation. We also found the number of tokens can be reduced via random masking. This indicates that ViToK compresses videos non-uniformly, with significant content condensed in a small number of tokens.

Future Work

Token length is a key factor for efficient media generation. The compression rate of modern auto-encoders is still less efficient than conventional video codec. For example, Movie Gen TAE compresses a 10s 768p video to ~300K tokens with a size of ~10M bytes. This is about 10x of video codec like mp4. In our study of ViToK, we found that transformer-based 1D tokenization can encode videos non-uniformly, leading to a much smaller token length.

Following this direction, we can explore training generative models with tokens with varying lengths. This will significantly improve the efficiency of video generation. With 10x compression rate, the efficiency will improve 100x for quadratic-complexity models. With length-varying tokens, the generation can be at different levels of details, adaptive to use cases with different hardware, networking, and cost.

Systematically, we can combine the aforementioned technologies to build very efficient generative models. With that, it will be much easier to work with other explorations, e.g. Multimodal LLMs conditioning, RL finetuning, distillation, and downstream applications.