Yefei He

Mobile/Wechat：18888922934 | Email：billhe@zju.edu.cn | Homepage: https://hexy.tech/

EDUCATION

Zhejiang University *Ph.D. of Instrument Science (*Top 10%*)* *Sep 2021 - now*

Zhejiang University *Bachelor of Instrument Science (*Top 5% *)* *Sep 2017 - Jun 2021*

PROFESSIONAL EXPERIENCE

Shanghai AI Laboratory *Research Intern* *Apr2024 - Mar2025*

- **Research on efficient vision-language models:** designing efficient inference framework for vision-language models with adaptive token-level sparse attention and KV cache compression.
- **Research on efficient visual generation models:** improving the inference efficiency of Lumina-mGPT and Lumina-mGPT 2.0 models with quantization; designing a parallel decoding framework (ZipAR) for AR visual generation models; designing a new AR visual generation paradigm (NAR), outperforming VAR on both performance and efficiency.

Tongyi Lab, Alibaba Group *Research Intern* *Apr2025 - now*

- **Research on post-training of Wan’s T2I/T2V models:** building high-quality data curation and filtering pipeline; training high performance reward models and performing RL to improve Wan’s text rendering capability.
- **Research on unified understanding and generation multimodal models:** designing hybrid textual & visual chain-of-thoughts process; designing 1D visual tokenizer with high-level semantics and low-level details.

PUBLICATIONS

Mover: Motion-controllable Video Generation via Latent Trajectory Guidance *Under Review*

Ruihang Chu, Yefei He*, Zhekai Chen, Shiwei Zhang*

Abstract: We present Mover, a simple and scalable framework that brings motion control to video generative models. Our core idea is to directly make the original condition features motion-aware for guiding video synthesis. To this end, we first represent object motions with dense point trajectories, allowing fine-grained control over the scene. We then project these trajectories into latent space and propagate the first frame's features along each trajectory, producing an aligned spatiotemporal feature map that is naturally integrated into the off-the-shelf image-to-video model. Through scaled training, Mover generates 5-second, 480p videos whose motion controllability rivals Kling 1.5 Pro's commercial Motion Brush.

Neighboring Autoregressive Modeling for Efficient Visual Generation. *ICCV 2025*

Yefei He, Yuanyu He, Shaoxuan He, Feng Chen, Hong Zhou, Kaipeng Zhang, Bohan Zhuang

Abstract: We propose Neighboring Autoregressive Modeling (NAR), a novel paradigm that formulates autoregressive visual generation as a progressive outpainting procedure, following a near-to-far “next-neighbor prediction” mechanism. To enable parallel prediction of multiple adjacent tokens in the spatial-temporal space, we introduce a set of dimension-oriented decoding heads, each predicting the next token along a mutually orthogonal dimension. Experiments on ImageNet 256×256 and UCF101 demonstrate that NAR achieves 2.4× and 8.6× higher throughput respectively, while obtaining superior FID/FVD scores for both image and video generation.

ZipAR: Accelerating Autoregressive Image Generation through Spatial Locality. *ICML 2025*

Yefei He, Feng Chen, Yuanyu He, Shaoxuan He, Hong Zhou, Kaipeng Zhang, Bohan Zhuang

Abstract: We propose a training-free, plug-and-play parallel decoding framework for accelerating autoregressive (AR) visual generation. The motivation stems from the observation that images exhibit local structures, and spatially distant regions tend to have minimal interdependence. Given a partially decoded set of visual tokens, in addition to the original next-token prediction scheme in the row dimension, the tokens corresponding to spatially adjacent regions in the column dimension can be decoded in parallel. Experiments demonstrate that ZipAR can reduce the number of model forward passes by up to 91% on the Emu3-Gen model without requiring any additional retraining.

ZipVL: Efficient large vision-language models with dynamic token sparsification. *ICCV 2025*

Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, Bohan Zhuang

Abstract: We present an efficient inference framework designed for LVLMs through a dynamic ratio allocation strategy of important tokens. This ratio is adaptively determined based on the layer-specific distribution of attention scores, rather than fixed hyper-parameters, thereby improving efficiency for less complex tasks while maintaining high performance for more challenging ones. Then we select important tokens based on their normalized attention scores and perform sparse attention mechanism solely on those important tokens, reducing the latency in the prefill phase. Tokens deemed less important will be discarded to reduce KV cache size, alleviating the memory bottleneck in the decoding phase.

ZipCache: Accurate and Efficient KV Cache Quantization with Salient Token Identification *NeurIPS 2024*

Yefei He, Luoming Zhang, Weijia Wu, Jing Liu, Hong Zhou, Bohan Zhuang

Abstract: We present an accurate and efficient KV cache quantization method for large language models (LLMs). We propose normalized attention score as an effective metric for identifying salient tokens by considering the lower triangle characteristics of the attention matrix. The quantization bit-width for each token is then adaptively assigned based on their saliency. Moreover, we develop an efficient approximation method that decouples the saliency metric from full attention scores, enabling compatibility with fast attention implementations like FlashAttention. When evaluating Mistral-7B model on GSM8k dataset, ZipCache is capable of compressing the KV cache by 4.98×, with only a 0.38% drop in accuracy.

EfficientDM: Efficient Quantization-Aware Fine-Tuning of Low-Bit Diffusion Models *ICLR 2024*

Yefei He, Jing Liu, Weijia Wu, Hong Zhou, Bohan Zhuang

Abstract: We introduce a data-free, quantization-aware and parameter-efficient fine-tuning framework for low-bit diffusion models, dubbed EfficientDM, to achieve QAT-level performance with PTQ-like efficiency. Specifically, we propose a quantization-aware variant of the low-rank adapter (QALoRA) that can be merged with model weights and jointly quantized to low bit-width. The fine-tuning process distills the denoising capabilities of the full-precision model into its quantized counterpart, eliminating the requirement for training data. We also introduce scale-aware LoRA optimization and temporal learned step-size quantization to further enhance performance. There is only a marginal 0.05 sFID increase when quantizing both weights and activations of LDM-4 to 4-bit on ImageNet 256×256.

PTQD: Accurate Post-Training Quantization for Diffusion Models *NeurIPS 2023*

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, Bohan Zhuang

Abstract: We propose a unified formulation for the quantization noise and diffusion perturbed noise in the quantized denoising process. We first disentangle the quantization noise into its correlated and residual uncorrelated parts regarding its full-precision counterpart. The correlated part can be easily corrected by estimating the correlation coefficient. For the uncorrelated part, we subtract the bias from the quantized results to correct the mean deviation and calibrate the denoising variance schedule to absorb the excess variance resulting from quantization. We also introduce a mixed-precision scheme for selecting the optimal bitwidth for each denoising step.

BiViT: Extremely Compressed Binary Vision Transformers *ICCV 2023*

Yefei He, Zhenyu Lou, Luoming Zhang, Jing Liu, Weijia Wu, Hong Zhou, Bohan Zhuang

Abstract: we have proposed to tackle two fundamental challenges with customized solutions for BiViTs, and have successfully applied BiViTs to visual tasks for the first time. To deal with the long-tailed distribution of softmax attention, we have proposed the Softmax-aware Binarization for self-attention. To preserve information from pretrained model, we have proposed the Cross-layer Binarization scheme that decouples the quantization of self-attention and MLPs. Moreover, we have introduced Parameterized Weight Scales to enhance representational ability of BiViTs.

HONORS

Graduate National Scholarship *2024*

Graduate National Scholarship *2022*

First-class Scholarship of Zhejiang University *2020*