MD-DiT: Step-aware Mixture-of-Depths for Efficient Diffusion Transformers

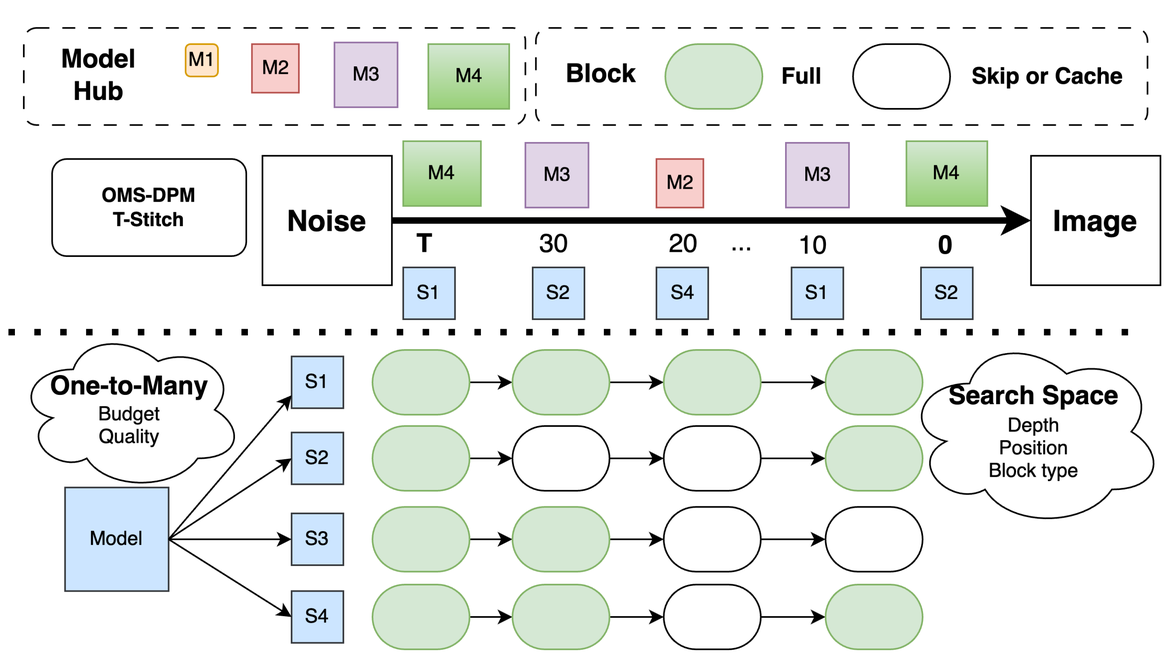
Mingzhu Shen[1*], Pengtao Chen[2*], Peng Ye[34], Guoxuan Xia[1], Tao Chen[2], Christos-Savvas Bouganis[1], Yiren Zhao[1]
[1]Imperial College London, [2]Fuda n University, [3]CUHK, [4]Shanghai AI Lab, [*]Equal Contribution

Framework

MD-DiT, a **one-to-many** unified framework that realizes a **mixture-of-depths across different timesteps** via the incorporation of block **skipping and caching** techniques.

Our contributions are as follows:

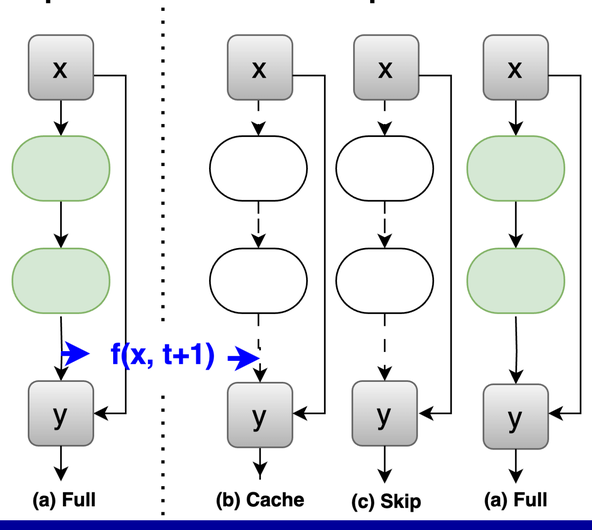
- We introduce MD-DiT, a framework that combines block skipping and caching to create a **mixture-of-depths across timesteps.**
- We explore **depth allocation strategies** for each timestep and use a **gradient-free search method** to identify a more compact model, improving diffusion transformer acceleration.



Allocating different computations to each timestep

Search Space

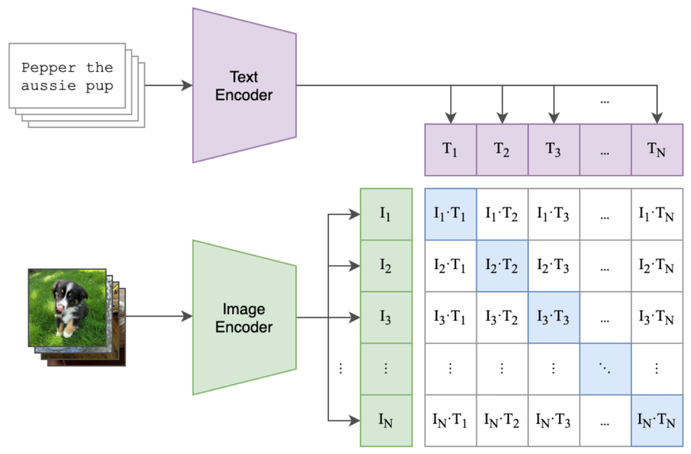
Search to Skip or Cache or Compute



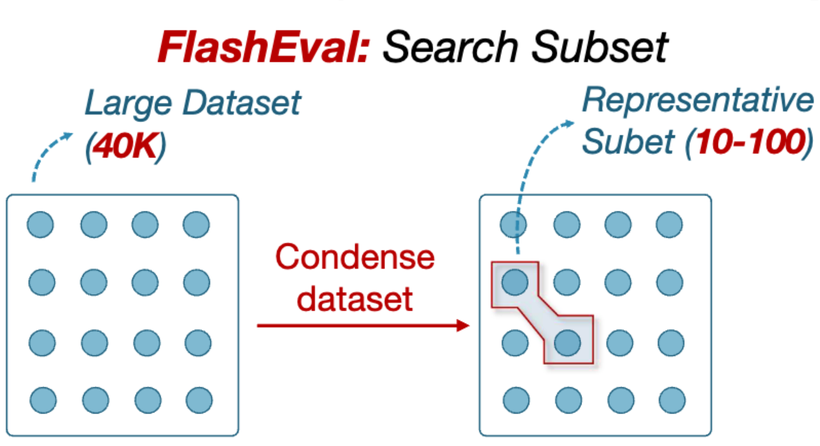
Search Insights

How to Search Efficiently : **Clip Score** on a small **Search SubSet**

Gradient-Free Search: Covariance Matrix Adaptive Evolution Strategy



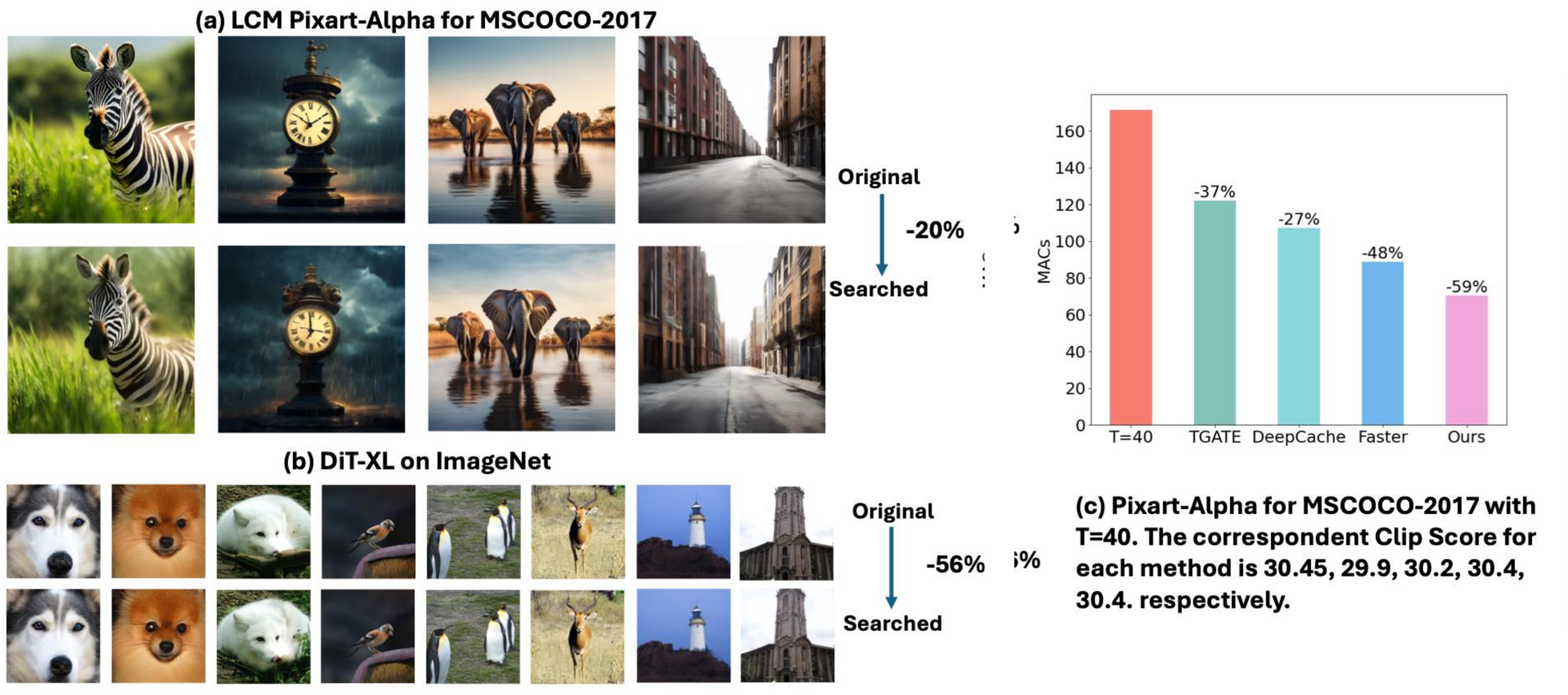
CLIPScore: A Reference-free Evaluation Metric for Image Captioning



FlashEval: Towards Fast and Accurate Evaluation of Text-to-image Diffusion Generative Models

Experiments

Through extensive experiments, we have successfully compressed the **LCM-4Step** model with a **20% reduction** in Multiple-Accumulate Operations (MACs). This achievement is further amplified in a **40-step** setting, where we have accomplished a **59% reduction**.



(c) Pixart-Alpha for MSCOCO-2017 with T=40. The correspondent Clip Score for each method is 30.45, 29.9, 30.2, 30.4, 30.4. respectively.