EC-DiT

Scaling Diffusion Transformers with Adaptive Expert-Choice Routing

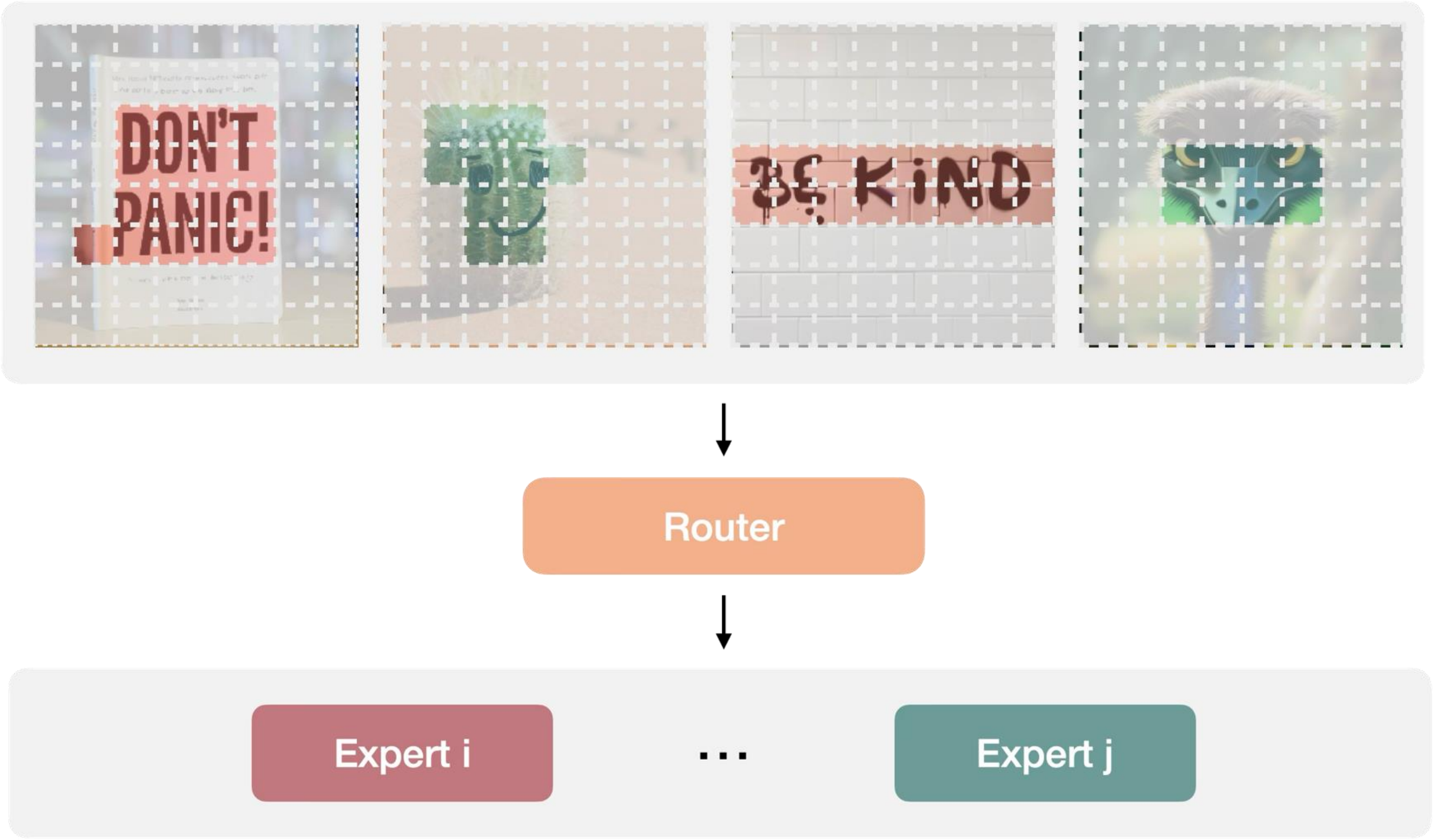
Haotian Sun [1], Tao Lei[2], Bowen Zhang[2], Yanghao Li[2], Haoshuo Huang[2], Ruoming Pang[2], Bo Dai [1], Nan Du[2]

[1] Georgia Institute of Technology
[2]Apple

PAPER

Motivation

Global batch information



Tokens are processed

non-autogresssively

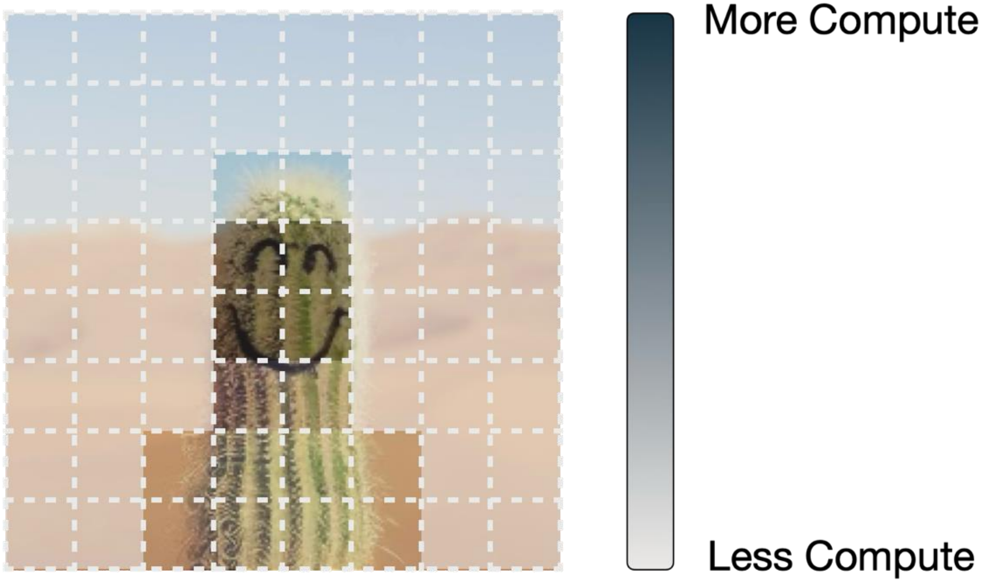
Router sees every token

Effective routing

Motivation

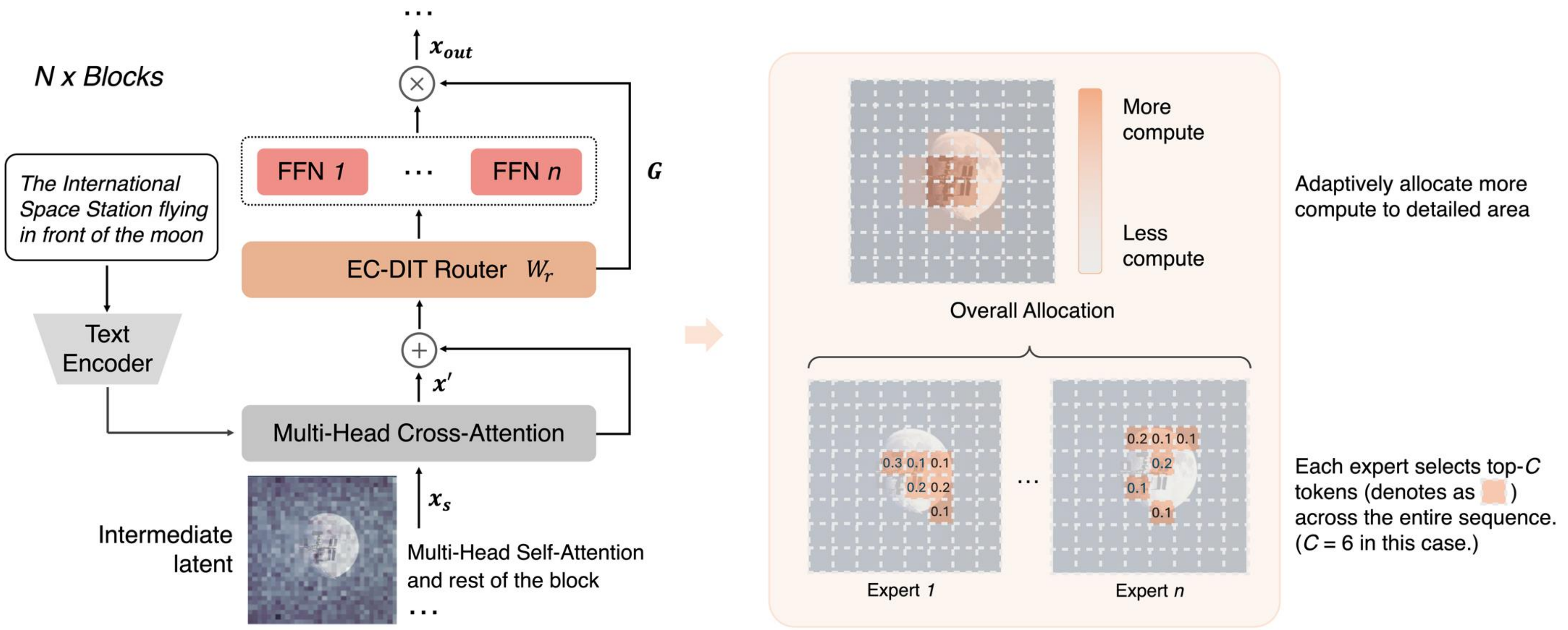
Heterogenous compute allocation

"A small cactus with a happy face in the Sahara desert"



Expert-choice routing

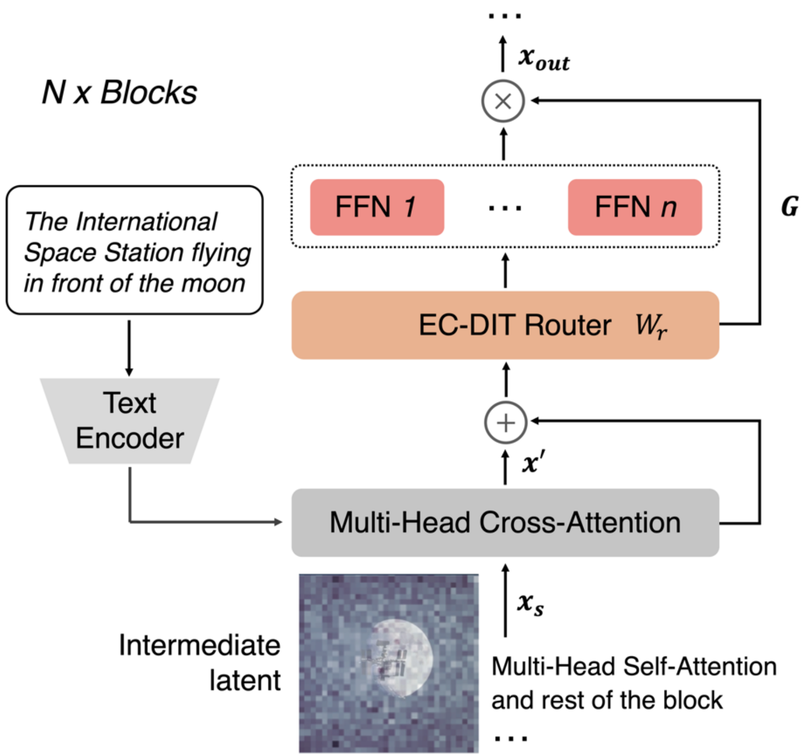
Heterogenous compute allocation



EC-DiT leverages sequence-wide information to route tokens to experts adaptively. This dynamic routing allocates more computation to detailed areas (like the space station and moon) while reducing it for simpler regions like the background.

Expert-choice routing

Heterogenous compute allocation



Algorithm 1 Pseudocode of EC-DiT's Routing Layer

```
# B: batch size, S: sequence length, d: hidden dimension
# E: number of experts, C: expert capacity
# experts: list of length E containing expert FFNs
def ec_dit_routing(x_p, W_r, experts):
    # 1. Compute token-expert affinity scores
    logits = einsum('bsd,de->bse', x_p, W_r)          # shape: (B, S, E)
    affinity = softmax(logits, dim=-1)                # shape: (B, S, E)
    affinity = einsum('bse->bes', affinity)           # shape: (B, E, S)
    # 2. Select the top-k tokens for each expert
    gating, index = top_k(affinity, k=C, dim=-1)      # shape: (B, E, C)
    dispatch = one_hot(index, num_classes=S)          # shape: (B, E, C, S)
    # 3. Process the tokens by each expert and combine
    x_in = einsum('becs,bsd->becd', dispatch, x_p)  # shape: (B, E, C, d)
    x_e = [experts[e](x_in[:, e]) for e in range(E)]
    x_e = stack(x_e, dim=1)                           # shape: (B, E, C, d)
    x_out = einsum('becs,bec,becd->bsd', dispatch, gating, x_e)
    return x_out                                      # shape: (B, S, d)
```

EC-DiT leverages sequence-wide information to route tokens to experts adaptively. This dynamic routing allocates more computation to detailed areas (like the space station and moon) while reducing it for simpler regions like the background.

Experiment Setup

Model configs & sizes

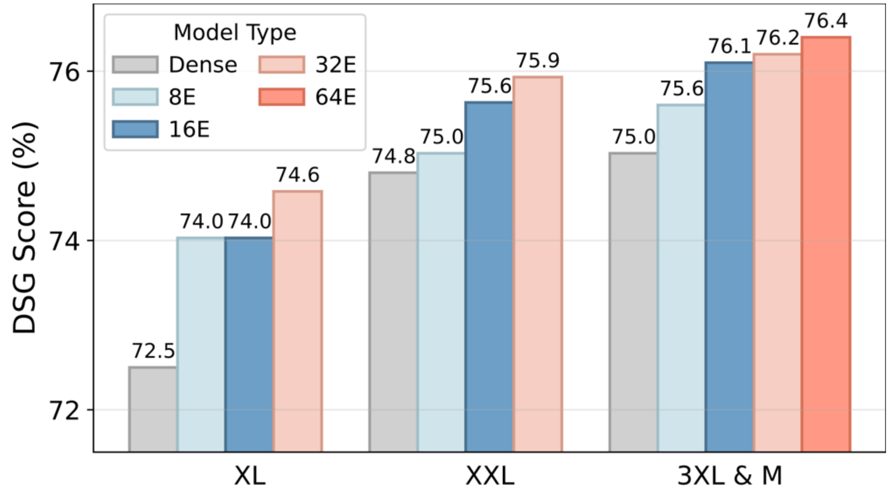
Config.	Total Params.					Activated Params.	Model Arch.			
	DENSE	8E	16E	32E	64E	EC-DIT	#Layers	Hidden dim.	#Head	#KV
XL	1.47B	2.51B	3.70B	6.08B	–	1.62B	28	1,152	18	6
XXL	2.35B	4.87B	7.73B	13.47B	–	2.71B	38	1,536	24	6
3XL	4.50B	10.74B	17.87B	32.15B	–	5.18B	42	2,304	36	6
M	8.03B	–	–	–	97.21B	8.27B	38/46*	3,072	48	12

We adopt a modified DiT architecture with additional cross-attention modules for text-to-image generation. We freeze a 670M clip-based text encoder with the T5 tokenizer and a 34M variational autoencoder (VAE) with 8 channels (CLIP-ViT-bigG). The transformer component is configured with 4 model sizes.

Generation Performance

T2I Alignment

Model (↓) / Score (%) (→)	Overall	Single obj.	Two obj.	Counting	Colors	Position	Color attr.
SD v1.5 (Rombach et al., 2022)	43.00	97.00	38.00	35.00	76.00	4.00	6.00
PixArt-α (Chen et al., 2023a)	48.00	98.00	50.00	44.00	80.00	8.00	7.00
SD v2.1 (Rombach et al., 2022)	50.00	98.00	51.00	44.00	85.00	7.00	17.00
DALL-E 2 (Ramesh et al., 2022)	52.00	94.00	66.00	49.00	77.00	10.00	19.00
SDXL (Podell et al., 2023)	55.00	98.00	74.00	39.00	85.00	15.00	23.00
SDXL Turbo (Podell et al., 2023)	55.00	**100.00**	72.00	49.00	80.00	10.00	18.00
IF-XL (Saharia et al., 2022)	61.00	97.00	74.00	66.00	81.00	13.00	35.00
DALL-E 3 (Shi et al., 2020)	67.00	96.00	87.00	47.00	83.00	**43.00**	45.00
SD3-Large (Esser et al., 2024)	68.00	98.00	84.00	66.00	74.00	40.00	43.00
SD3-Large (Esser et al., 2024) w/ DPO	71.00	98.00	**89.00**	73.00	83.00	34.00	47.00
DENSE-3XL	68.92	99.69	86.00	69.41	81.67	20.58	56.19
EC-DIT-3XL-32E	70.91	99.64	87.88	72.53	83.84	21.19	60.40
EC-DIT-M-64E	**71.68**	99.84	88.67	**73.69**	**85.77**	21.33	**60.80**



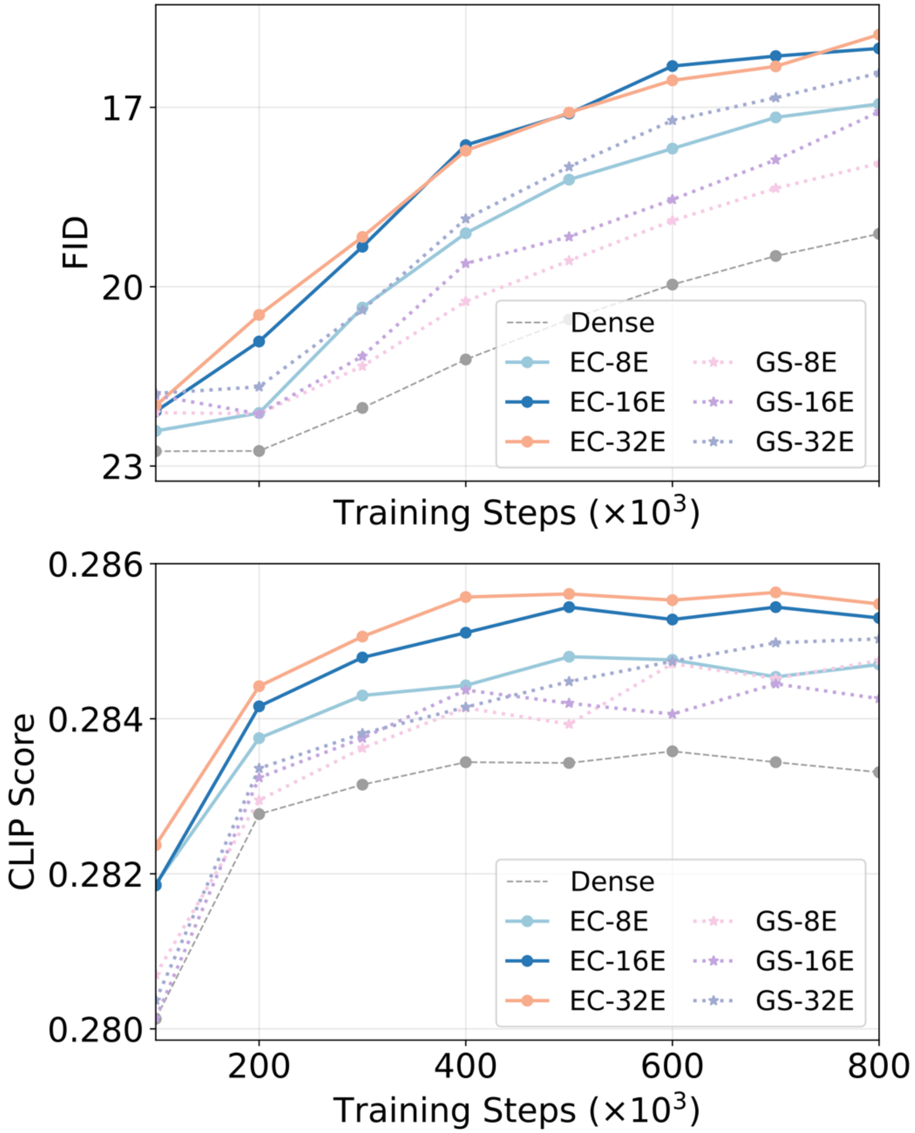
EC-DiT outperforms dense models w/ competitive inference speed.

Our largest model (64 experts) hits a GenEval and DSG of **71.68%** and **76.4%**, respectively, w/ around 23% additional overhead to the dense model.

Generation Performance

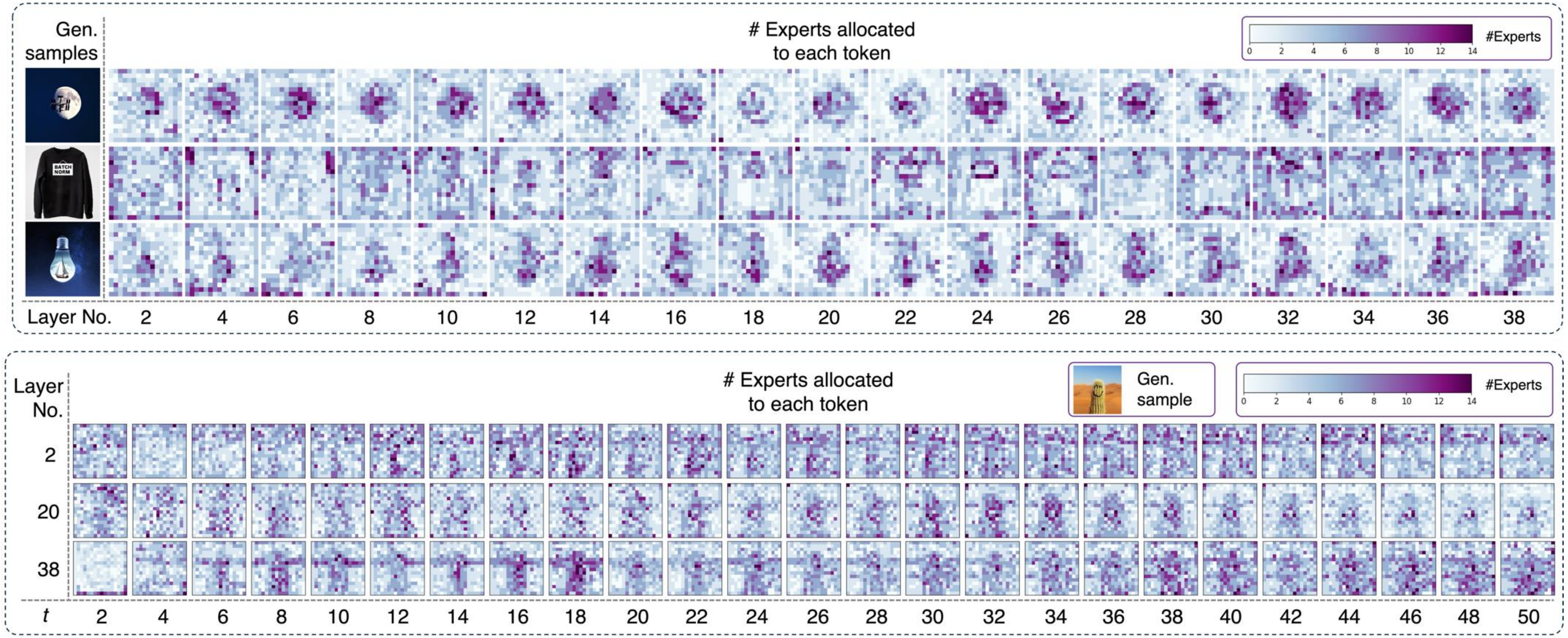
Outperforming over baselines

EC-DiT consistently shows better training convergence and performance. With 8 experts, it rivals the token-choice MoEs with 16 experts, and more experts lead to even more significant gains.



Visualization

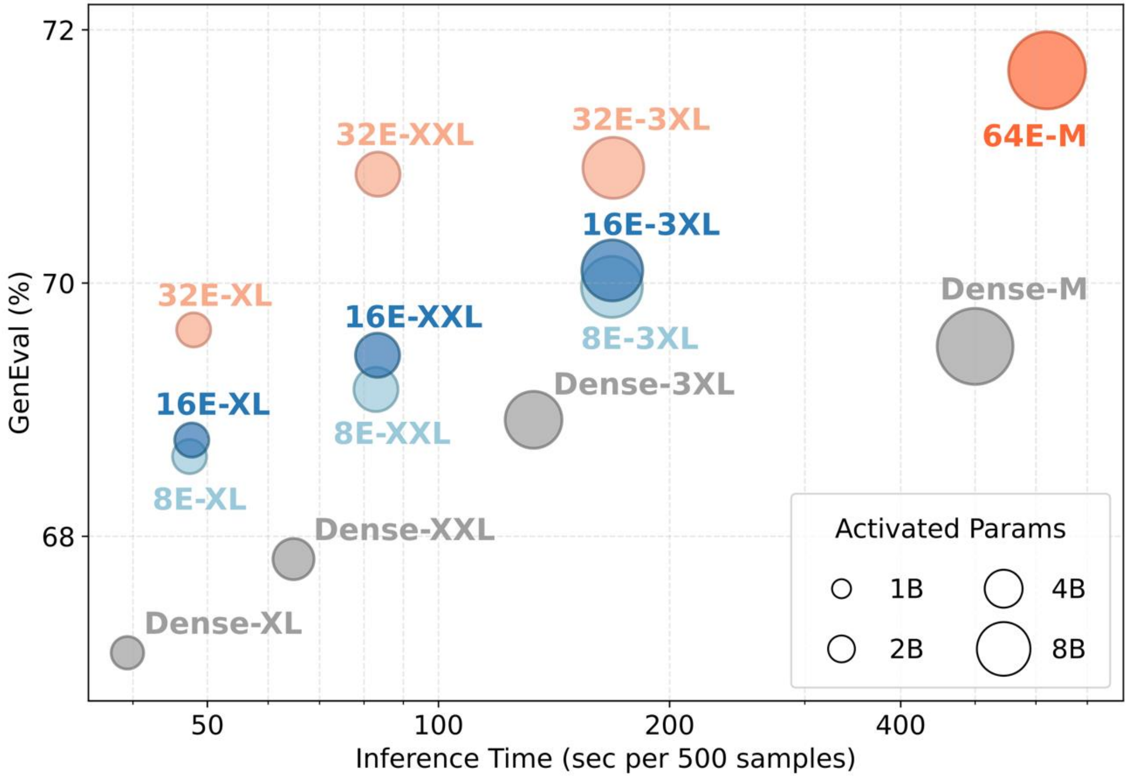
Heterogenous compute allocation



Darker areas in the heatmap highlight where more compute is focused—like the moon and rendered text—showing **EC-DiT** ’s awareness of textual importance and its ability to prioritize key elements during generation.

Scalability

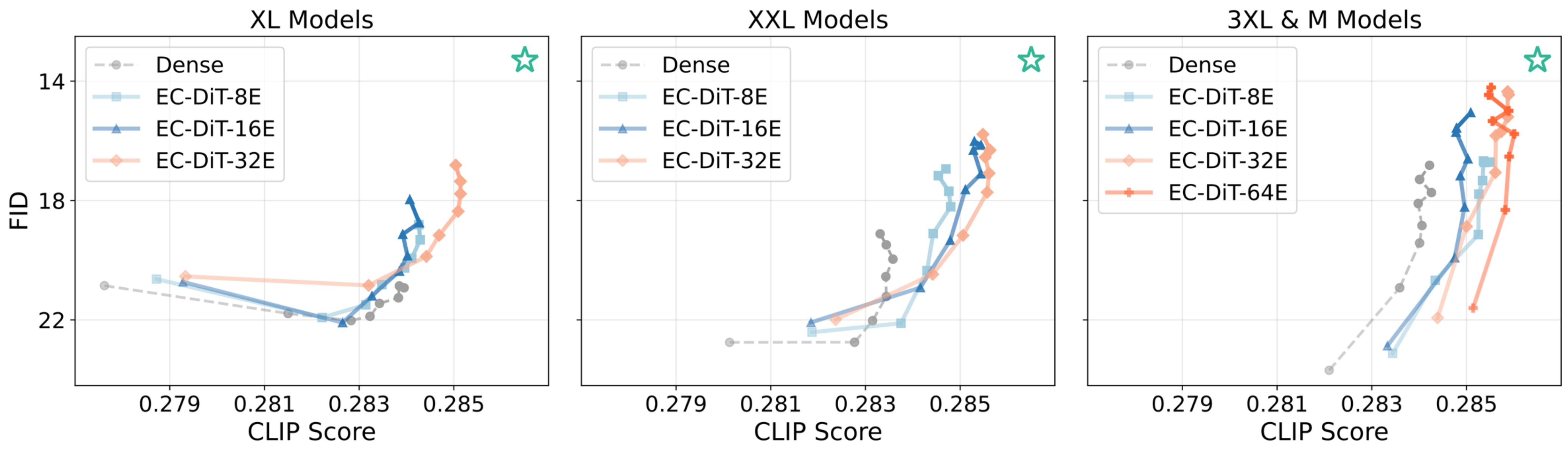
Inference Efficiency



EC-DiT shows superior performance compared to dense models, with **less than 30%** additional overhead. Note that the overhead difference from the theoretical number might be attributed to the varying efficiency in inference-time parallelism.

Scalability

FID & CLIP Score

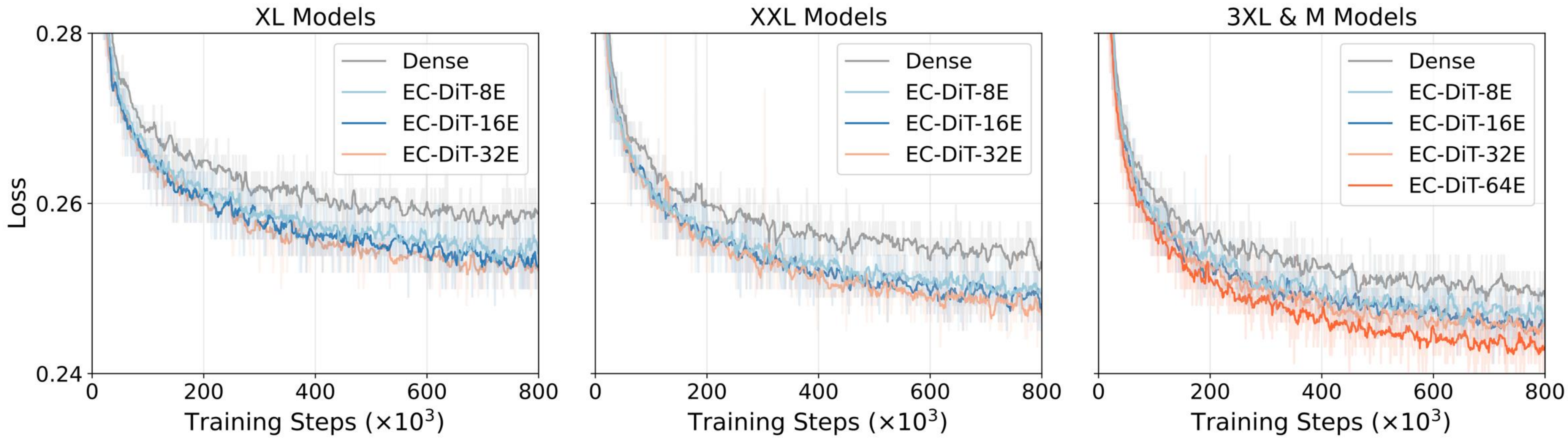


EC-DiT with more experts consistently yields performance gains in image generation quality and text-image alignment.

Scalability

Training Dynamics

EC-DiT brings a significant improvement in loss reduction over the dense models throughout the training period and across all model settings.



Thank you.

EC-DiT: Scaling Diffusion Transformers with Adaptive Expert-Choice Routing

PAPER