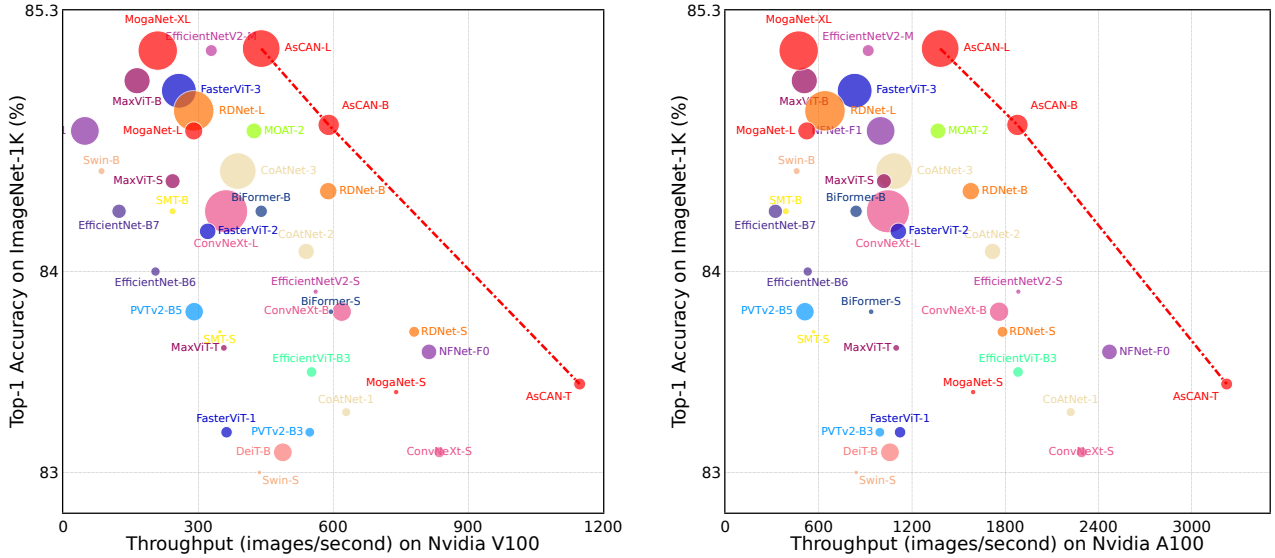

Figure 1: **ImageNet-1k task.** We plot the latency measured as throughput in images inferred per second on a single V100 GPU (*Left*)/A100 GPU (*Right*) with batch-size 16. The area of each circle is proportional to the number of model parameters. While some models regress between two hardware (*e.g.*, MaxViT-S vs SMT-B), our model consistently achieves better accuracy *vs* latency trade-offs.

Table 1: **ImageNet-1K Class Conditional Generation.** We train two variants of our T2I UNet architecture for conditional generation at 256 and 512 resolutions. Our architectures achieve better FID vs throughput trade-off than existing state-of-the-art architectures.

Model	256 × 256			512 × 512		
	FLOPs	Throughput A100 (B=64) samples/sec	FID	FLOPs	Throughput A100 (B=64) samples/sec	FID
LDM	104G	362	3.60	-	-	-
U-ViT-L/2	77G	498	3.40	340G	86	4.67
U-ViT-H/2	133G	271	2.29	546G	45	4.05
DiT-XL/2-G	118G	293	2.27	525G	51	3.04
Ours	52G	556	2.23	224G	130	3.15

Table 2: **C before T analysis.** We ablate over the preference of **C** and **T** blocks in a stage. Using **T** block in stem / early layers result in lower throughput. Further, the performance of configurations with **T** before **C** yields lower accuracy vs throughput trade-off.

Block Configuration	Params	Inference(images/s) A100 B=16	A100 B=64	V100 B=16	Top-1 Acc.
`CC-CCCT-CCTT-CTTT` (C1)	55M	3224	4295	1148	83.4%
`CC-CCCT-CCTT-CCTT` (C2)	73M	3217	4179	1036	83.2%
`CC-CCCT-CCTT-TTTT` (C3)	41M	3384	4472	1224	82.9%
`CC-CCCT-CCCC-TTTT` (C4)	50M	3434	4411	1182	83.1%
`CC-CCCT-CCCT-CCCT` (C5)	95M	3135	4066	991	82.7%
`CC-TCCC-CCTT-CTTT` (T1)	56M	3029	4021	945	82.8%
`CC-CCCT-TTCC-CTTT` (T2)	57M	3100	4092	1021	82.9%
`CC-CCCT-CCTT-TTTC` (T3)	64M	3190	4193	1045	83.1%
`TT-CCCT-CCTT-CTTT` (T4)	55M	1428	1584	487	83.1%
`TT-TTTC-TTCC-TTTC` (T5)	100M	1280	1440	390	83.5%

Table 3: **Results on ImageNet-1K for Classification Task.** We benchmark related works pointed by reviewers (see procedure in Appendix Table 7). We include memory consumed during inference with 64 batch-size.

	Architecture	Resolution	Params	MACs	Throughput (images/s) Batch (B) A100 B=1	A100 B=16	A100 B=64	V100 B=1	V100 B=16	Top-1 Accuracy	Memory (GB) B=64
ConvNet	RDNet-S	224	50M	8.7G	102	1782	2761	53	780	83.7%	12.6
	RDNet-B	224	87M	15.4G	80	1578	1891	48	589	84.4%	17.5
	RDNet-L	224	186M	34.7G	78	640	990	32	290	84.8%	26.2
Hybrid	MOAT-2	224	73M	17.2G	132	1367	2087	38	424	84.7%	31.8
	MOAT-3	224	190M	44.9G	70	805	962	21	246	85.3%	65.2
	SMT-S	224	21M	4.7G	38	566	2211	43	348	83.7%	13.5
	SMT-B	224	32M	7.7G	27	388	1412	14	243	84.3%	22.6
	MogaNet-S	224	25M	5.0G	93	1593	2455	47	740	83.4%	17.1
	MogaNet-L	224	83M	15.9G	34	523	882	24	290	84.7%	46.6
	MogaNet-XL	224	181M	34.5G	32	471	576	19	210	85.1%	74.9
	BiFormer-S	224	26M	4.5G	45	937	2139	36	595	83.8%	18.6
	BiFormer-B	224	57M	9.8G	50	840	1439	28	440	84.3%	27.9
	RMT-S	224	27M	4.5G	46	790	2439	38	480	84.1%	13.7
Ours	AsCAN-T	224	55M	7.7G	199	3224	4295	67	1148	83.44%	9.6
	AsCAN-B	224	98M	16.7G	113	1878	2393	38	590	84.73%	15.2
	AsCAN-L	224	173M	30.7G	120	1381	1617	40	440	85.24%	21.2