NanoSD: Edge Efficient Foundation Model for Real-Time Image Restoration

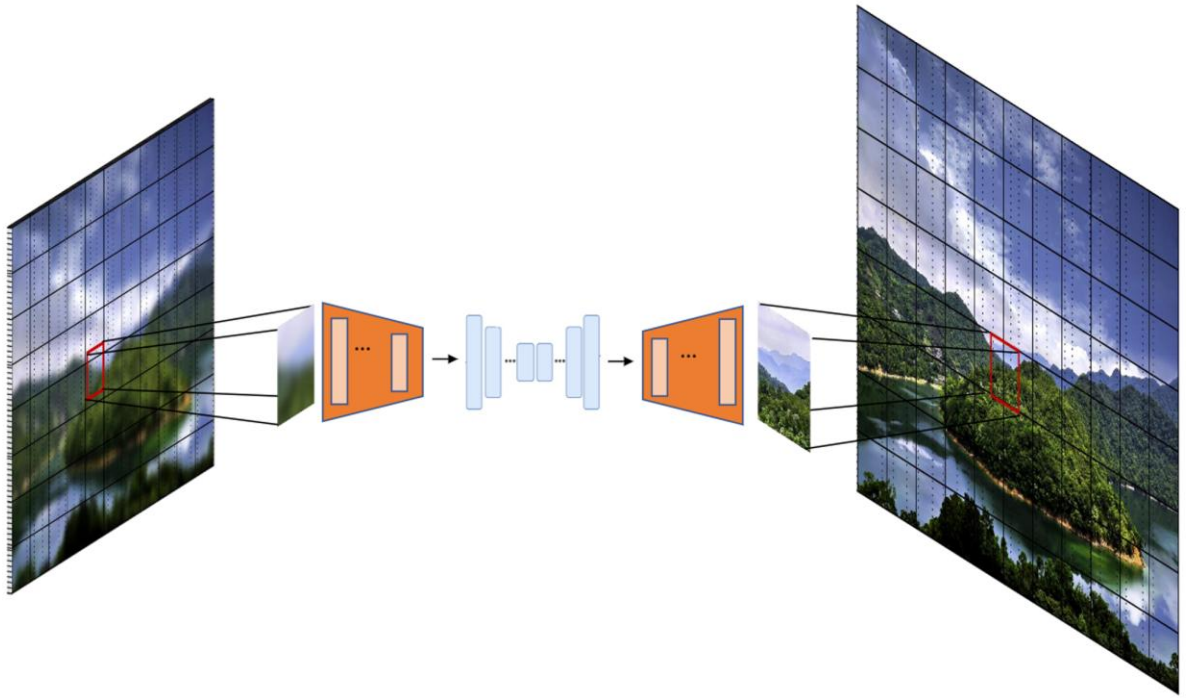
Subhajit Sanyal*, Srinivas Soumitri Miriyala* #, Akshay Janardan Bankar*, Manjunath Arveti, Sowmya Vajrala,

Shreyas Pandith, Sravanth Kodavanti,

Abhishek Ameta, Harshit, Amit Satish Unde # | Samsung Research Institute Bangalore, India | *Equal Contribution |

#Presenters

Background: Mobile Image Signal Processing using SD 1.5



- Mobile ISP necessitates tiling for Image Restoration (IR), where each tile is processed sequentially using the method for IR.

Memory Burden

- SD 1.5 = ~860M params = 0.8GB in INT8 (w8)

Latency Burden

- 1000 x 750 image → 88 tiles of 128 x 128
- SD 1.5 model in w4 quantization = 116ms
- End-to-End latency = 116 x 88 = 10.2 seconds per single time step of diffusion

Motivation: One-for-all Edge Efficient Vision Foundation Model

Stable Diffusion 1.5

✓ Strong generative prior

✗ Heavy U-Net + VAE

✗ Not practical for high-res image restoration on edge

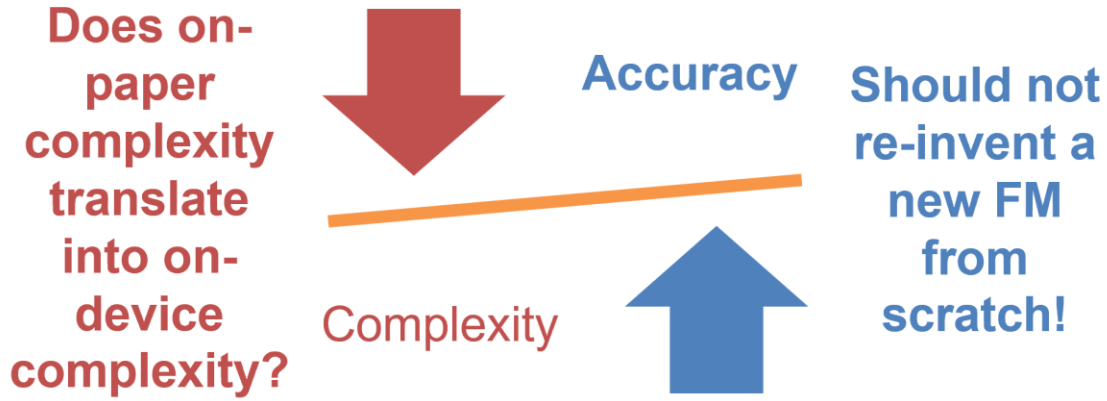
Existing lightweight diffusion models

✓ Faster than baseline SD

✗ Often task-specific

✗ May disrupt latent manifold / generative prior

✗ Limited support for diverse restoration plugins



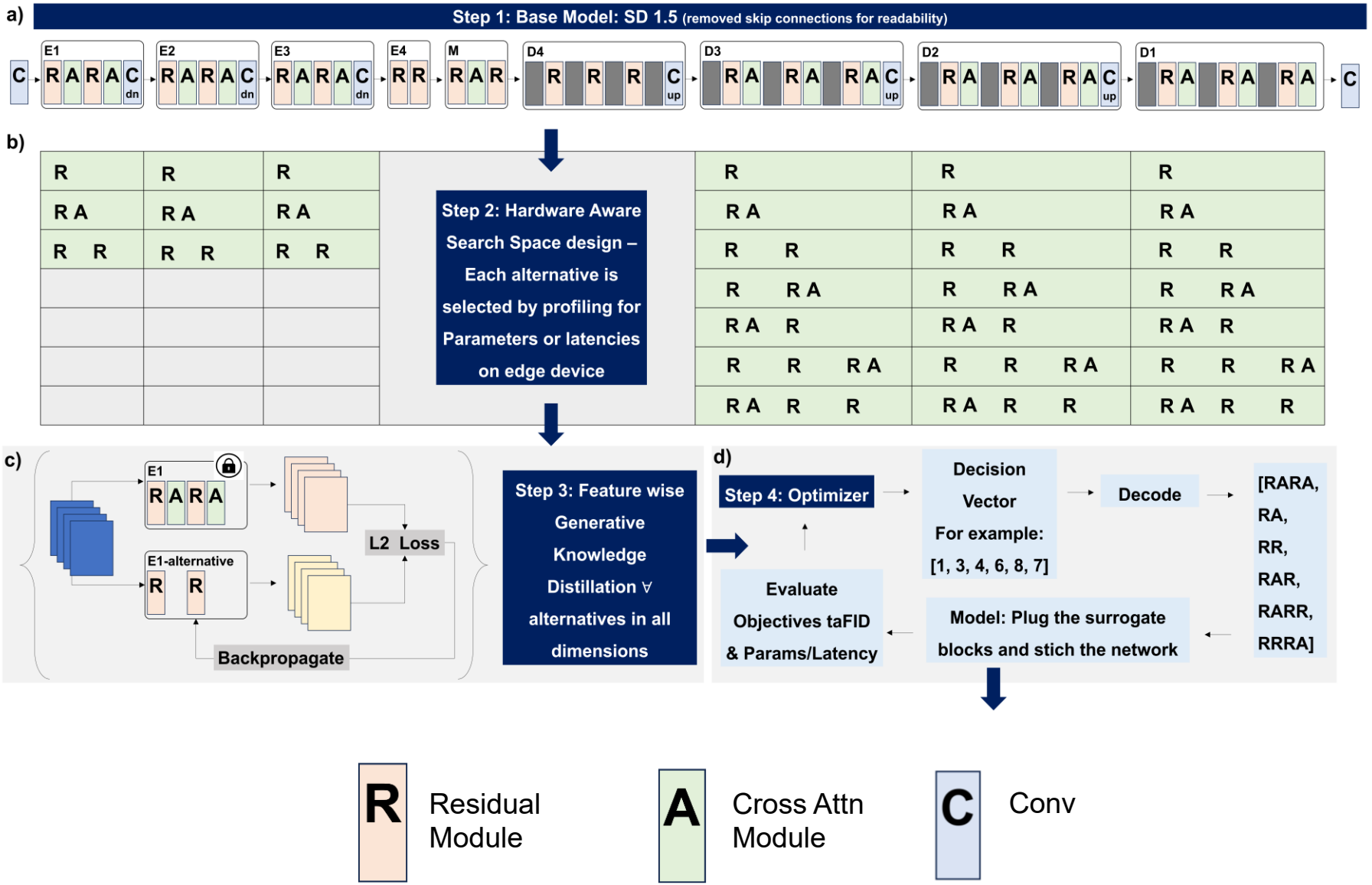
Key insight

- FLOPs and parameter count alone do not predict real NPU latency.
- On-device Speed depends on operator layout, tensor shape, and memory behavior.

Efficient Hardware Aware Search-1/2

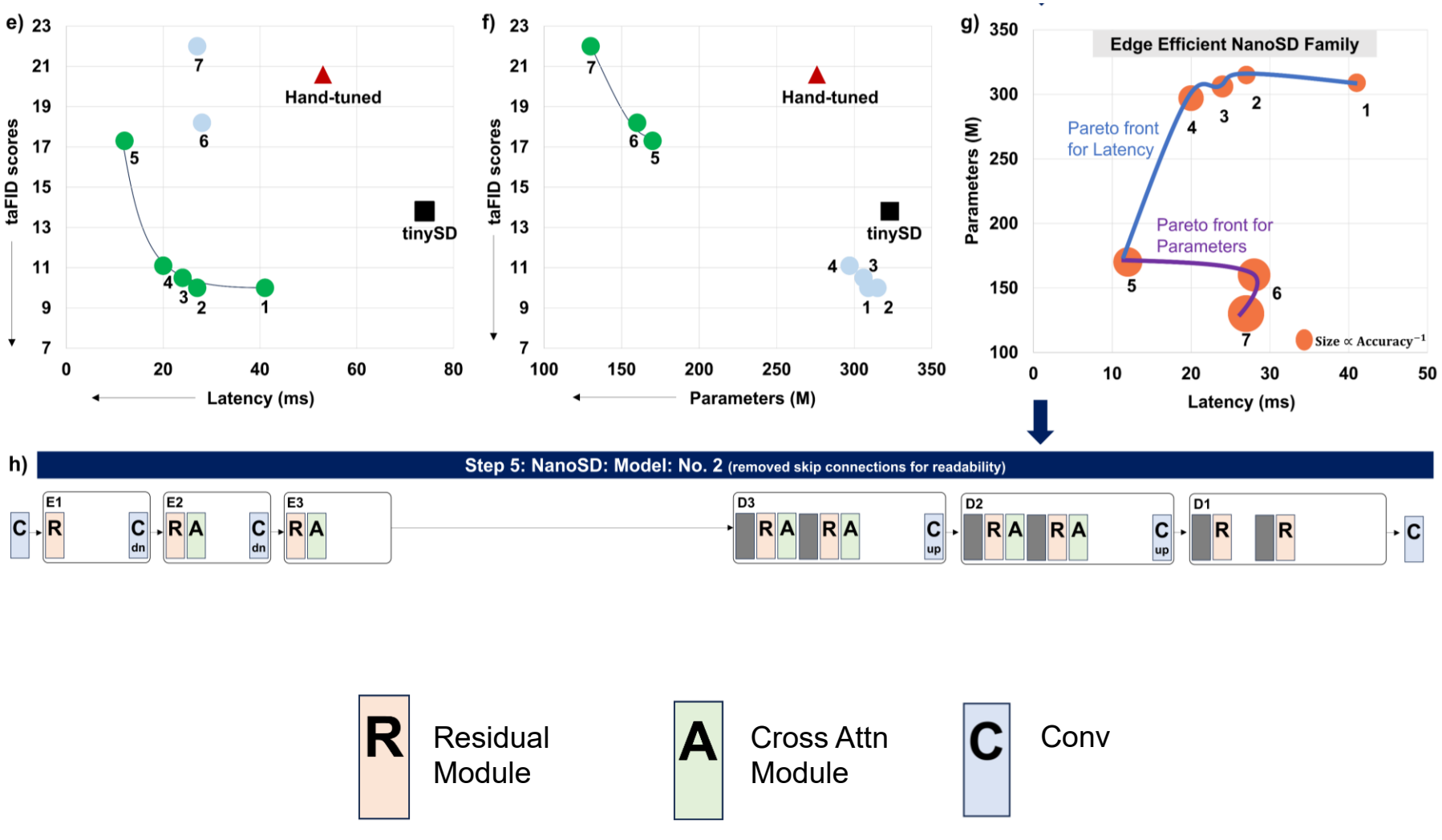
- For each base block (e.g. E2), the alternatives are designed by converting, quantizing, and profiling on HTP.
- FwKD is a shallow learning problem, drastically reducing the GPU infra needs – further only 25% of data needed during fwKD.
- After FwKD, alternatives are pruned if they are inaccurate and remaining turn into H/W-friendly surrogates to the base blocks.

Search Space Size = 32768
Surrogates Trained = 30
Decision Space = 6D, Objective = 2
INLP formulation – Plug & Play
Bayesian Opt, E2E training-free



Efficient Hardware Aware Search-2/2

- This approach results in Agile search, thereby allowing multiple iterations of search, each with a different direction, e.g., focus on latency, or focus on lightweight, or focus on increasing accuracy.
- Thus, the proposed approach allows for optimized network surgery on an existing SoTA with a specific focus on edge deployment without re-inventing the wheel of model development.



Results

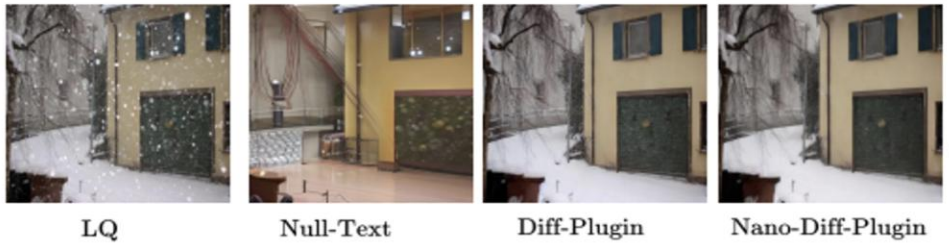
Model	Params (M)	tAFID	E1	E2	E3	E4	M	D4	D3	D2	D1	QC NPU (GS25 Ultra) *	Apple A17 pro ANE (iphone16)
SD 1.5	860	--	RARA	RARA	RARA	RR	RAR	RRR	RARARA	RARARA	RARARA	NA	NA
TinySD	323	13.8	RA	RA	RA	**Eliminated based on Automated Network Surgery & Sensitivity Analysis**			RA	RA	RA	74	192
Hand-tuned	276	20.6	RA	RA	RA				RAR	RAR	RAR	53	133
NanoSD 1	309	10	R	RA	RA				RARA	RRA	RRA	41	82
NanoSD 2	315	10	R	RA	RA				RARA	RARA	RR	27	38
NanoSD 3	306	10.5	R	RA	RA				RARA	RRA	RR	24	34
NanoSD 4	297	11.1	R	RA	RA				RARA	RR	RR	20	31
NanoSD 5	170	17.3	R	R	R				RA	RR	RR	12	20
NanoSD 6	160	18.2	R	R	R				RA	RA	RA	28	68
NanoSD 7	130	22	R	R	R				R	RA	RRA	27	66

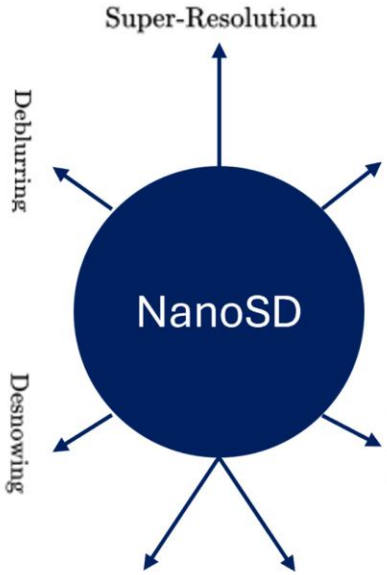
* Models are optimized for QC NPU & profiled on Apple ANE. The study merely shows that optimization on 1 hardware type extends to another type without any guarantee that the numbers are “optimized” for the other type

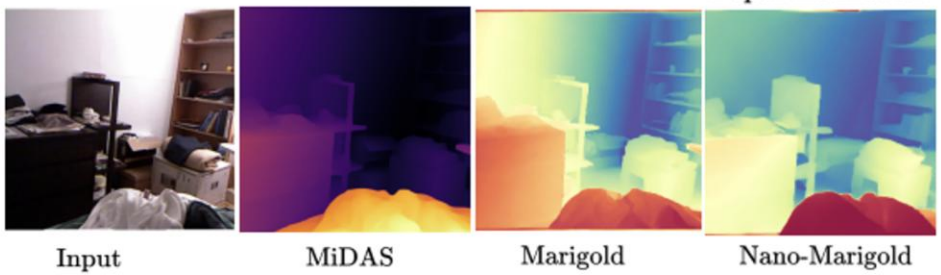
NanoSD: Edge Efficient Foundation Model for Real-Time Image Restoration

7 downstream vision use-cases

Preserves Generative Prior of SD 1.5 baseline







Takeaway: ~7x smaller edge-efficient SD foundation model for real-time 4k restoration under 2 seconds with generalization to 7 vision use-cases.

Thank You!