The University of Texas at Austin
Cockrell School of Engineering

Sept. 2025

CIRCUIT RESEARCH LAB (CRL) RESEARCH NOTE

Work for DAC - Design Automation Conference 2026

YIFAN WANG
Graduate Student, Fudan University
International Student, The University of Texas at Austin (Fall 2025)

PAPER REVIEW

PAPER REVIEW: “Diffusion Models” in DAC

TEXAS ENGINEERING

Left: What is the Diffusion model?

Right: What is U-net?

Z. Fan, S. Dai, R. Venkatesan, D. Sylvester and B. Khailany, "SQ-DM: Accelerating Diffusion Models with Aggressive Quantization and Temporal Sparsity," 2025 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2025, pp. 1-7, doi: 10.1109/DAC63849.2025.11132632.

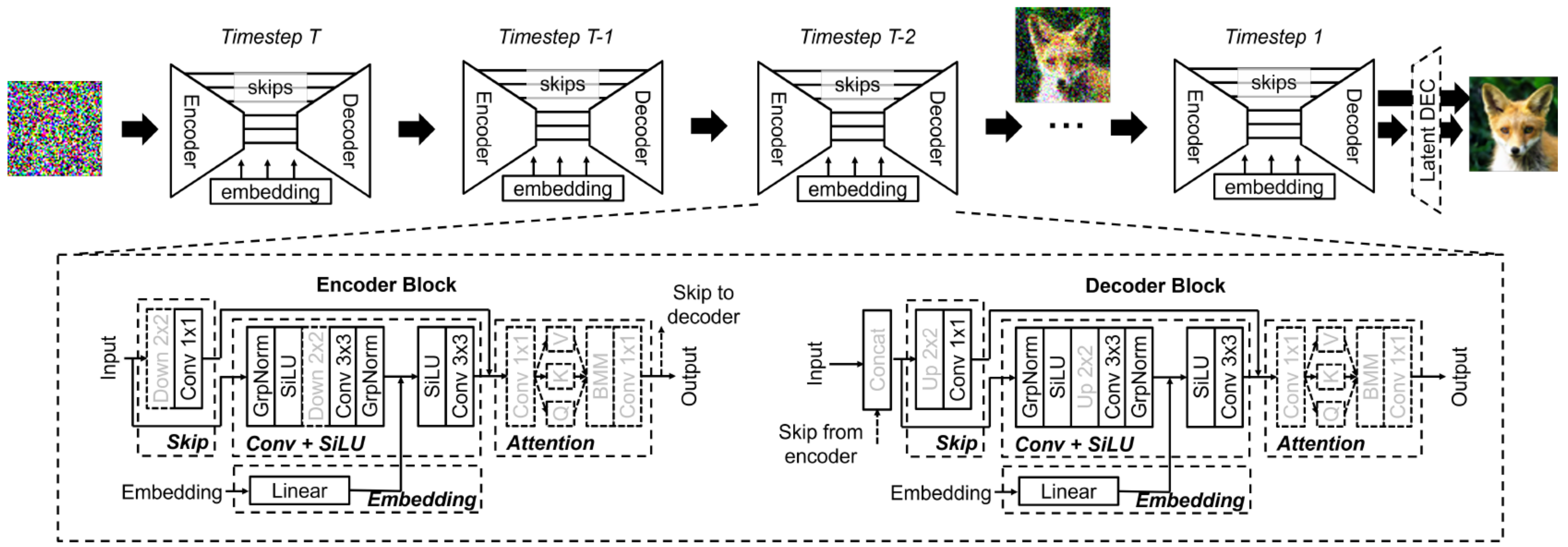
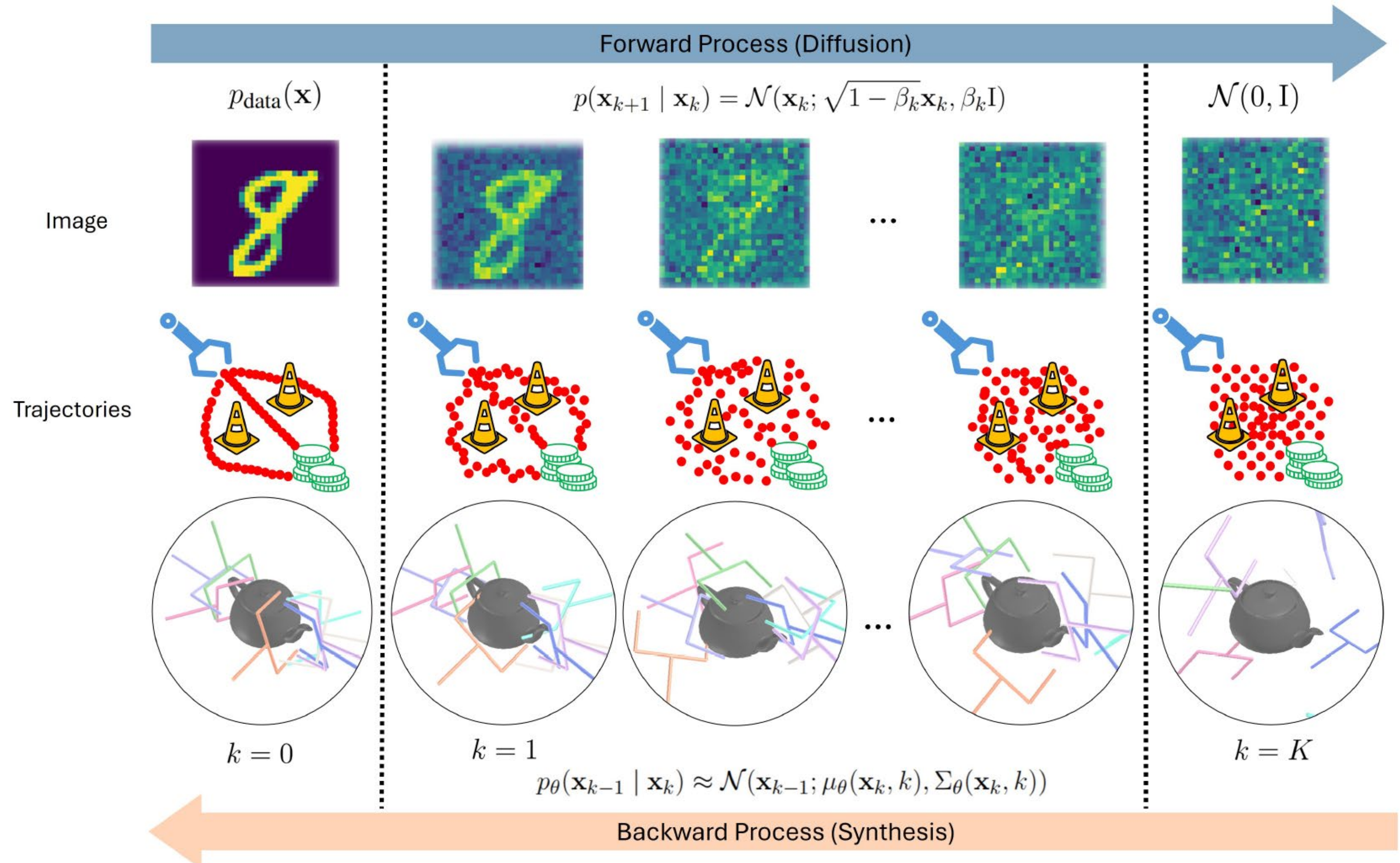


Fig. 2: Execution process and model architecture of EDM [4], [5].



Forward Diffusion

Reverse De-noise

Figure 1: Illustrations of diffusion (forward) processes on image, trajectories, and grasp poses (Urain et al. (2023)) and their corresponding synthesis (backward) processes.

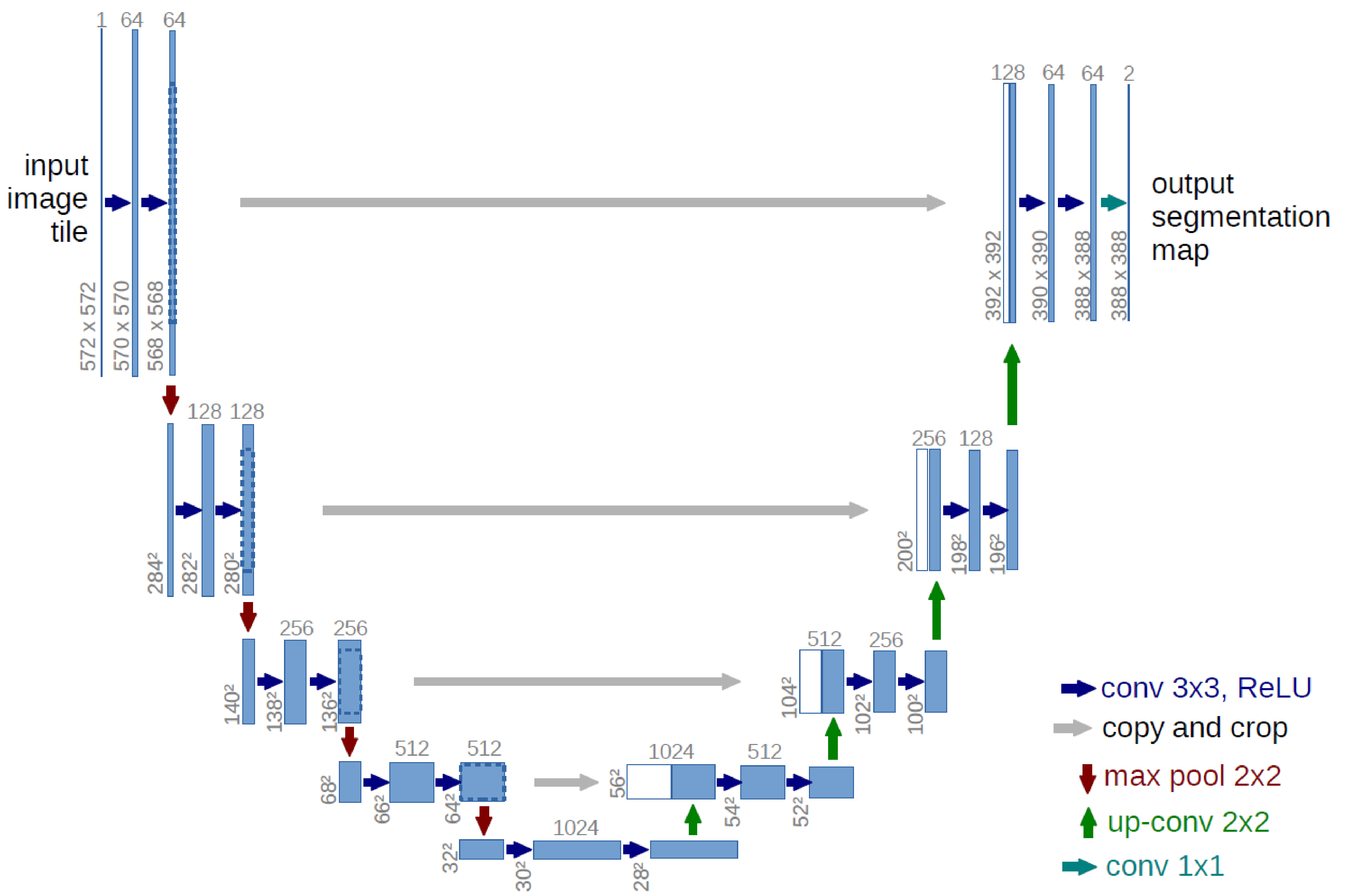


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-NET: Convolutional Networks for Biomedical Image Segmentation. In *Lecture notes in computer science* (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Wolf, R., Shi, Y., Liu, S., & Rayyes, R. (2025). Diffusion models for robotic manipulation: a survey. *Frontiers in Robotics and AI*, *12*. https://doi.org/10.3389/frobt.2025.1606247

PAPER REVIEW: “Diffusion Models” in DAC

TEXAS ENGINEERING

Recent work of Diffusion Models in DAC (mostly related to EDA field)

- Problem Based (mainly EDA field) Accelerator

① F. Azevedo, N. Lourenço and R. Martins, "Late Breaking Results: Encoder-Decoder Generative Diffusion Transformer Towards Push-Button Analog IC Sizing," 2025 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2025, pp. 1-2, doi: 10.1109/DAC63849.2025.11133224.

② X. Zheng, H. Gu, K. Peng, Y. Wang, W. Zhu and Z. Zhu, "Late Breaking Results: Customized Diffusion Model Empowered by Heterogeneous Graph Network for Effective Floorplanning," 2025 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2025, pp. 1-2, doi: 10.1109/DAC63849.2025.11133070.

③ Z. Wang et al., "DiffPattern: Layout Pattern Generation via Discrete Diffusion," 2023 60th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2023, pp. 1-6, doi: 10.1109/DAC56929.2023.10248009.

④ P. Haghi et al., "DM-Tune: Quantizing Diffusion Models with Mixture-of-Gaussian Guided Noise Tuning," 2025 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2025, pp. 1-7, doi: 10.1109/DAC63849.2025.11132501.

- Other

① Z. Fan, S. Dai, R. Venkatesan, D. Sylvester and B. Khailany, "SQ-DM: Accelerating Diffusion Models with Aggressive Quantization and Temporal Sparsity," 2025 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2025, pp. 1-7, doi: 10.1109/DAC63849.2025.11132632.

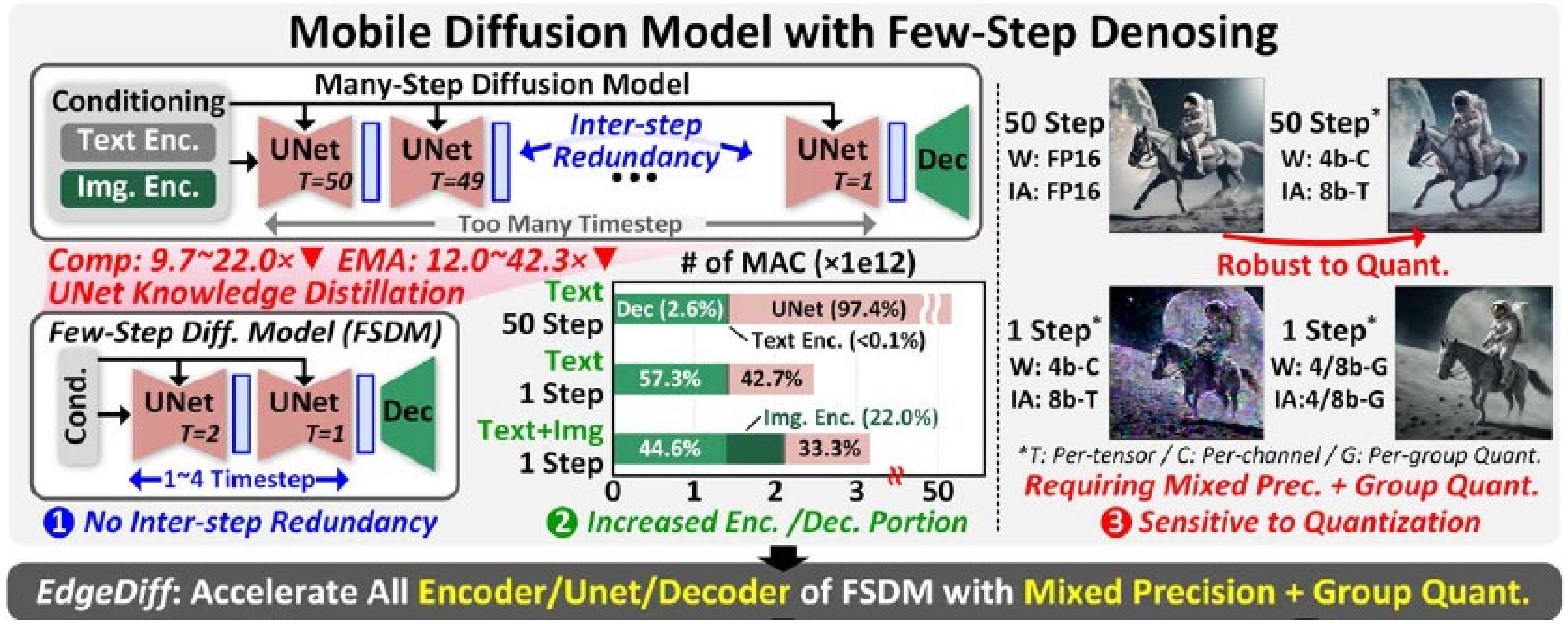
② Y. Park, S. Kim, Y. Kim, G. Ji and S. Ryu, "RADiT: Redundancy-Aware Diffusion Transformer Acceleration Leveraging Timestep Similarity," 2025 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2025, pp. 1-7, doi: 10.1109/DAC63849.2025.11133190.

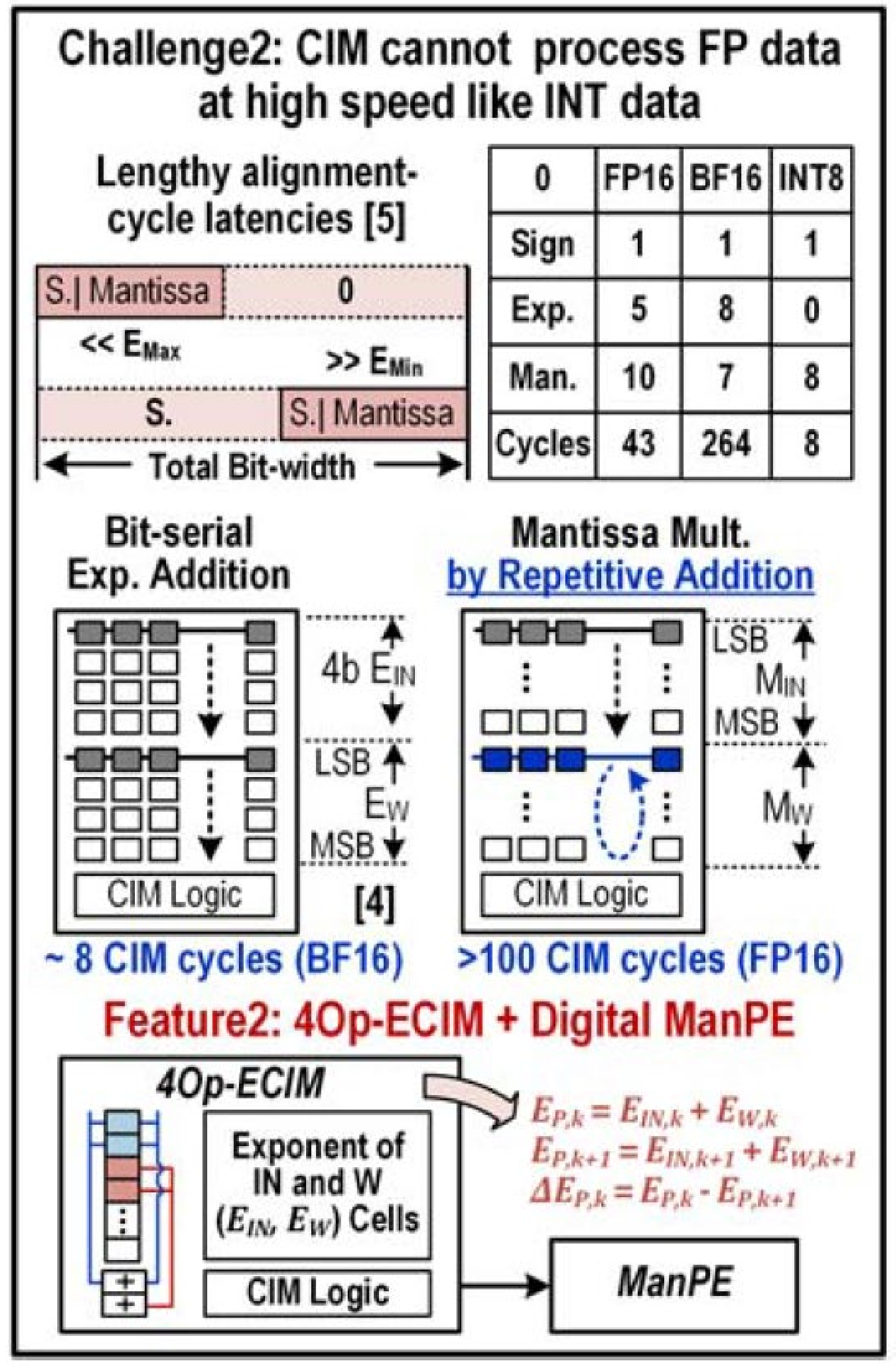
③ C. Qi et al., "MHDiff: Memory- and Hardware-Efficient Diffusion Acceleration via Focal Pixel Aware Quantization," 2025 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2025, pp. 1-7, doi: 10.1109/DAC63849.2025.11133171.

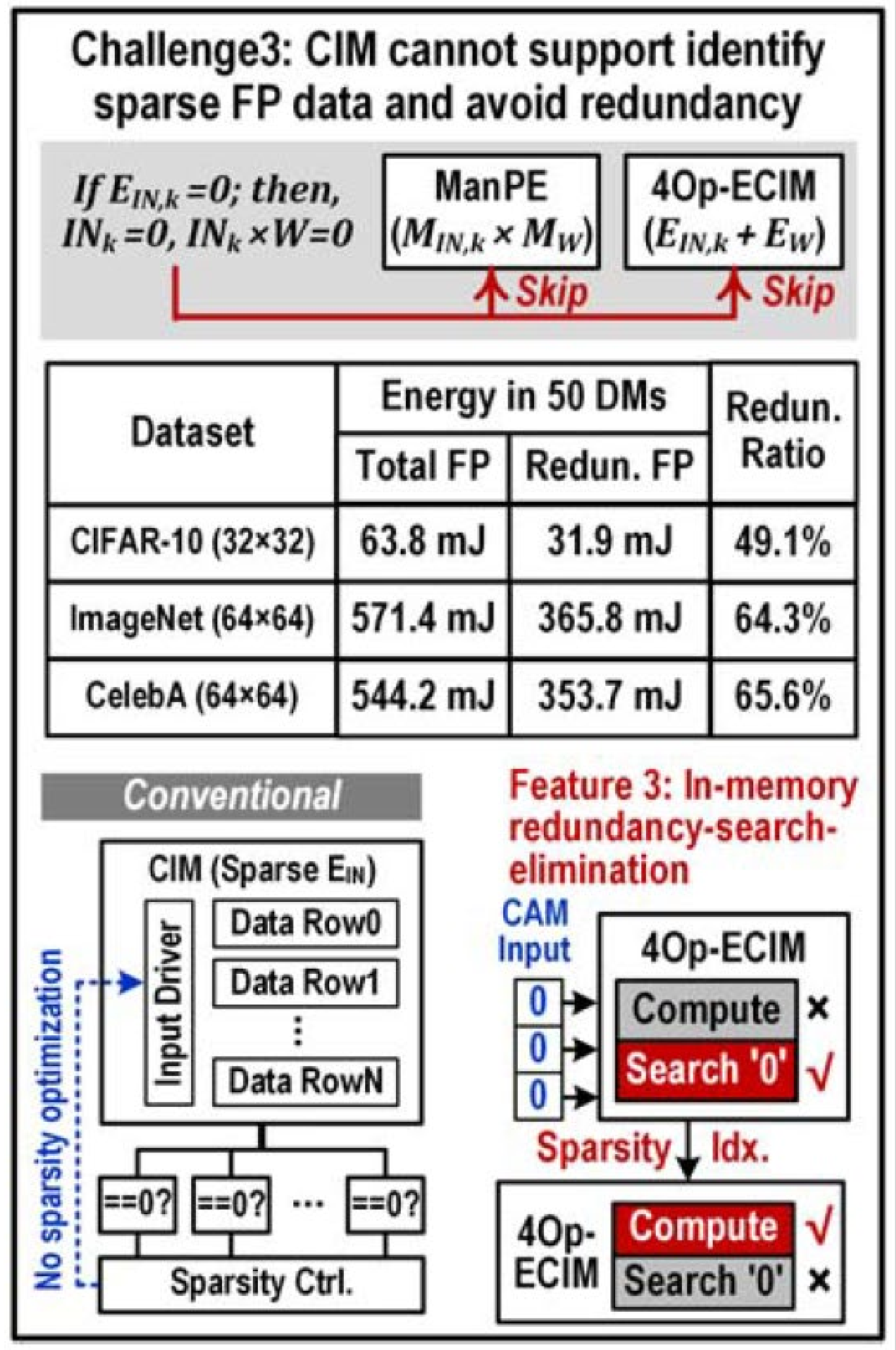
PAPER REVIEW: What they haven't done

TEXAS ENGINEERING

- What they did are (1) Quant (2) Sparsity (3) CIM-FP (4) Redundancy Detection







- What they may haven't done is (1) HW-optimization for 3D Diff

PAPER REVIEW: What they haven't done

TEXAS ENGINEERING

- What they may haven't done is (1) HW-optimization for 3D Diff

  (2) CIM technique is not common in DAC

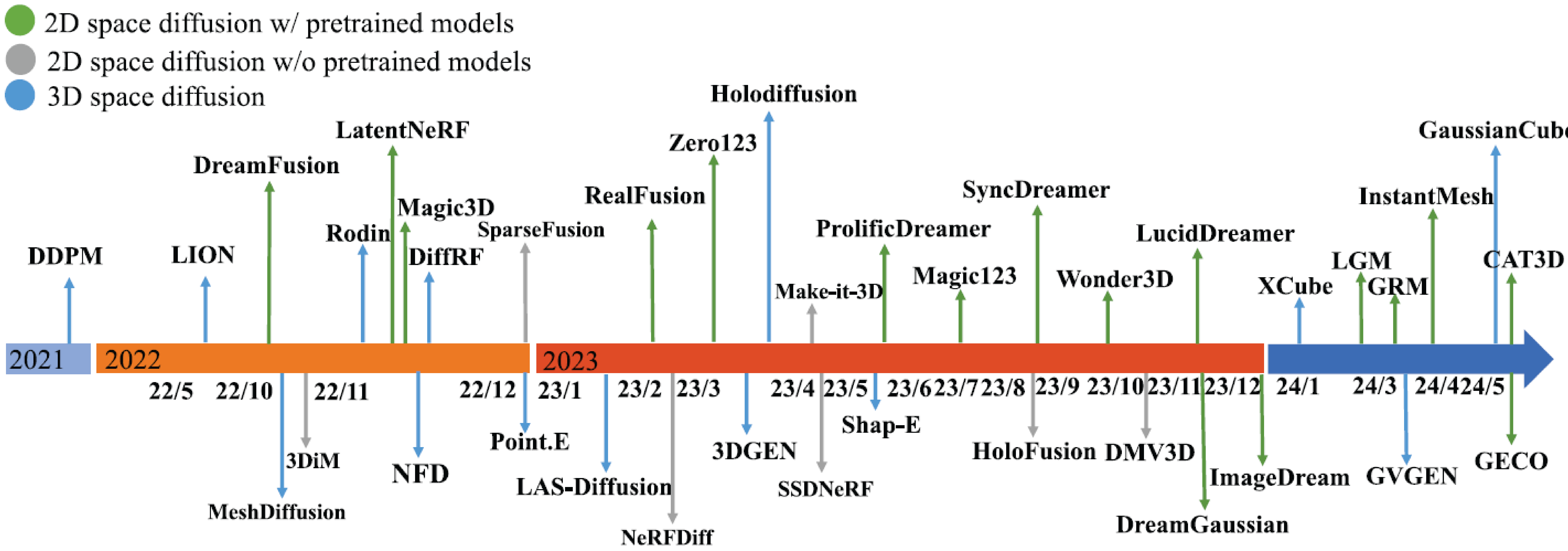
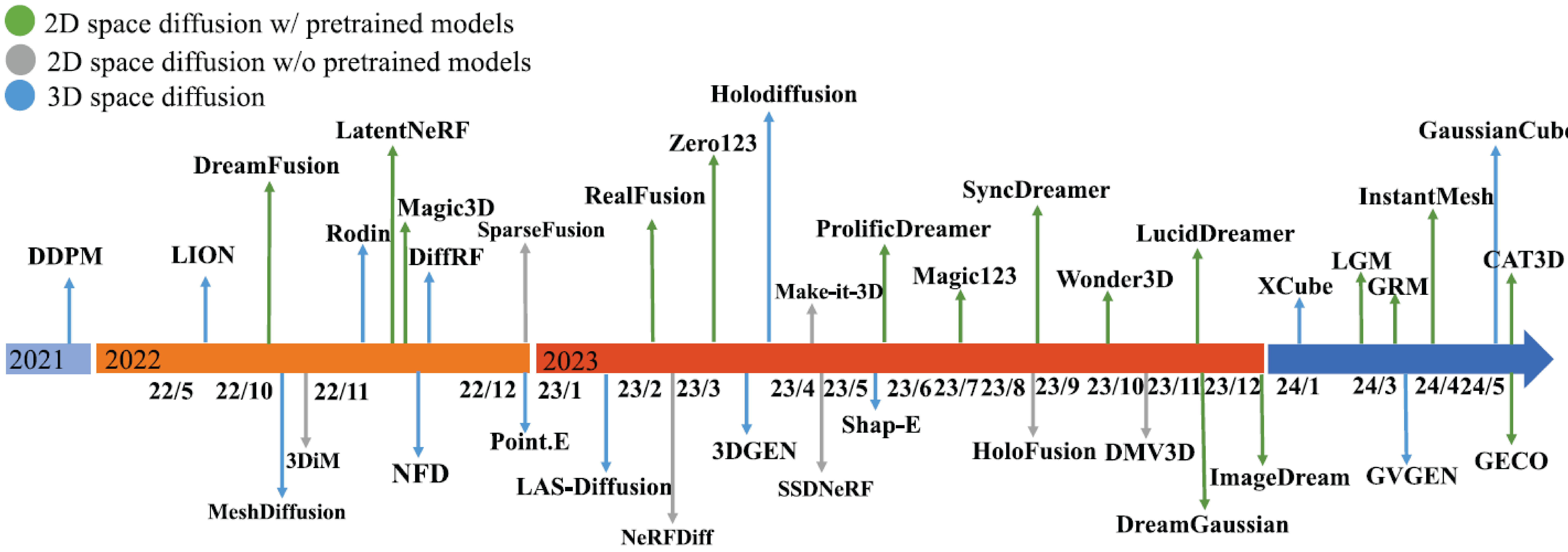


Fig. 1 A timeline of diffusion methods for 3D generation.

Diffusion models for 3D generation: A survey

PAPER REVIEW: What they haven't done

- Algorithm to HW? (1) The perpendicular gradient prevents the negative prompt from influencing the semantics of the positive prompt and makes the generation better conditioned on the prompts. (Text-to-3D) (2) Since a high-resolution SDF grid is both memory and computationally expensive, LAS-Diffusion uses a two-stage diffusion network: the first stage generates a low-resolution occupancy field to approximate the rough shape and the second stage generates detailed SDF values inside the occupied region. (3D diffusion using implicit representation)



Diffusion models for 3D generation: A survey

Fig. 1 A timeline of diffusion methods for 3D generation.

PAPER REVIEW: What they haven't done

- Area Specific? Robotic Manipulation

- Two main points must be considered to apply DMs to robotic manipulation.

- Firstly, in the diffusion processes described in the previous sections, given the initial noise, samples are generated solely based on the trained noise prediction network or conditional score network. However, robot actions are usually **dependent on simulated or real-world observations with multi-modal sensory data and the robot's proprioception**. Thus, the network used in the denoising process has to be conditioned on these observations.

- Secondly, unlike in image generation, where the pixels are spatially correlated, in trajectory generation for robotic manipulation, the **samples of a trajectory are temporally correlated**. On the one hand, generating complete trajectories may not only lead to high inaccuracies and error accumulation of the long-horizon predictions, but also prevent the model from reacting to changes in the environment. On the other hand, predicting the trajectory one action at a time increases the compounding error effect and may lead to frequent switches between modes.

PAPER REVIEW: Overall

TEXAS ENGINEERING

Article	Title	Year	Model	Solved Problem	Hardware Design	Contribution	HW-params	Others
DAC	DiffPattern: Layout Pattern Generation via Discrete Diffusion	2023	Discrete Diffusion Model	Generate Different Patterns from Single Topology.	no	develop a novel layout pattern generation method based on discrete denoising for synthesizing layout topology		
	DM-Tune: Quantizing Diffusion Models with Mixture-of-Gaussian Guided Noise Tuning	2025	DINOv2-ViT (feature extractor)	Generated images	no	(1) a fine-grained mixed-precision strategy can surpass full-precision models (2) integer-based strategies produce lower quality images. (3) 16-bit quantization offers a significant speedup with comparable performance to 32-bit and BF16 generally outperforms FP16 (4) For low-precision, FP8 (E4M3) is preferred over FP8 (E5M2) for better performance		
	Efficient Continuous Logic Optimization with Diffusion Model	2025	QoR surrogate model+diffusion model	logic synthesis optimization	no	proposed method not only achieves lower area and delay but also improves efficiency by 5X to 130X		
	Late Breaking Results: A Diffusion-Based Framework for Configurable and Realistic Multi-Storage Trace Generation	2025	diffusion-based synthetic trace generation framework	EDA synthesis	no	first work that utilizes the diffusion technique for the storage trace generation		
	Late Breaking Results: A Geometric Diffusion Model for Macro Placement Generation	2025	MacroDiff,	Macro placement	no	reduces macro overlap by 91.6%		
	Late Breaking Results: Customized Diffusion Model Empowered by Heterogeneous Graph Network for Effective Floorplanning	2025	heterogeneous graph convolutional network (HGCN) and graph attention blocks	floorplans	no	customized diffusion model to directly generate high-quality initial floorplans		
	Late Breaking Results: Encoder-Decoder Generative Diffusion Transformer Towards Push-Button Analog IC Sizing	2025	diffusion models (DMs) with an attention-based encoder-decoder	sizing of analog circuits	no	presenting higher generalization capabilities to performance targets not seen during training		
	MHDiff: Memory- and Hardware-Efficient Diffusion Acceleration via Focal Pixel Aware Quantization	2025	MHDiff	pixel-adaptive quantization	yes	(1) release memory burden (2) high- and low-precision quantization for focal pixels others respectively (3) use a PE array to process the condensed high-precision data through minor modifications to the PE units.	28nm, 500MHz, simwork	compared with SOTA: Cambricon-D
	RADiT: Redundancy-Aware Diffusion Transformer Acceleration Leveraging Timestep Similarity	2025	redundancy-aware DiT (RADiT)	DiT	yes	(1) identify data redundancy by evaluating blockwise input features and skip redundant computations by reusing results from consecutive timesteps (2) Dynamic Threshold Scaling Module (DTSM) and Compress and Compare Unit (CCU) are employed	28nm, 500MHz, simwork	compared with vanilla DiT hardware baseline
	SQ-DM: Accelerating Diffusion Models with Aggressive Quantization and Temporal Sparsity	2025	Elucidated Diffusion Models, EDM [4] and EDM2 [5], as our baseline diffusion models	CIFAR-10 AFHQv2 FFHQ ImageNet	yes	Our 4-bit quantization technique demonstrates superior generation quality compared to existing 4-bit methods	28nm, simwork	In the future, we plan to extend our techniques to diffusion models targeting video generation [39] and apply our methodology to other generative models.
ISSCC	An On-Device Generative AI Focused Neural Processing Unit in 4nm Flagship Mobile SoC with Fan-Out Wafer-Level Package	2025	An On-Device Generative AI Focused Neural Processing Unit in 4nm Flagship Mobile SoC with Fan-Out Wafer-Level Package	/	yes	We report on neural processing unit (NPU) in 4nm Samsung Exynos™ 2400 that employs heterogeneous architecture consisting of vector engines and two types of tensor engines	chipwork	Samsung Electronics, Hwaseong, Korea
	A 28nm 74.34TFLOPS/W BF16 Heterogenous CIM-Based Accelerator Exploiting Denoising-Similarity for Diffusion Models	2024	Diffusion Model	/	yes	(1) CIM with 2-bit parallel (2) FP: OP_CIM; Man_Digital (3) in memory redundancy search	28nm, BF16/FP16, chipwork	
	EdgeDiff: 418.4mJ/Inference Multi-Modal Few-Step Diffusion						28nm,	

The University of Texas at Austin
Cockrell School of Engineering