Ai4 | 2026

Shortcutting Time and Depth in Diffusion Models for Faster Agentic Applications

Denis Gudovskiy
Panasonic AI Lab
Mountain View, CA

Introduction

Denis Gudovskiy
Distinguished AI Engineer

- Joined Panasonic AI Lab in 2016
- 10+ years of experience developing AI/ML
- Collaborating with top research universities
- Leading AI development across multiple offices

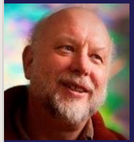
Efficient AI

THE UNIVERSITY OF ARIZONA

Prof: Huanrui Young
- Efficient AI Models to Reduce Costs
- System HW/SW Codesign
- Optimization of Agentic Applications

AI Agents

Berkeley UNIVERSITY OF CALIFORNIA



Prof: Kurt Keutzer
- Fast Multimodal AI Models
- Trustworthy UI Agents
- Synopsys ex-CTO, 30+ startups

1. Context

Diffusion Models in Agentic Applications

A case for augmenting or replacing autoregressive (AR) LLMs

Why Diffusion Models?

Diffusion models augment or replace autoregressive (AR) LLMs in many data domains:

- Fast parallel token generation vs. sequential in autoregressive models
- Built-in error correction and quality control via iterative denoising
- Native support of multimodality with continuous and discrete tokens

Image Generation'22

Stable Diffusion

Video Generation'24

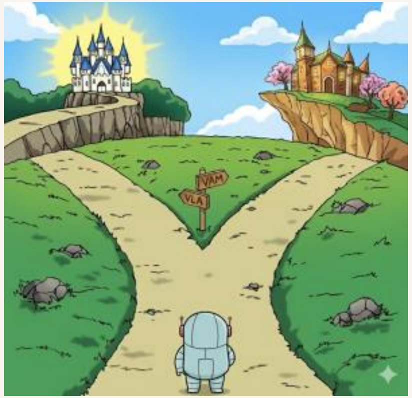
SORA

Text Generation'25

Discrete Diffusion
dLLMs Inception Labs

Robotic Control'24/26

Physical Intelligence's VLA
NVIDIA's world-action model

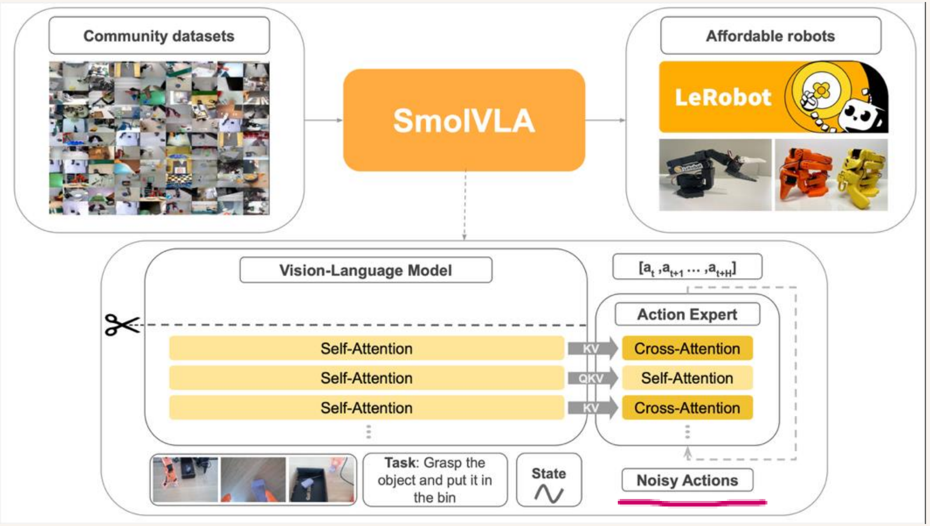


VLAs for Robotic Applications

Example: what is a Vision-Language Action (**VLA**) model:

1. Pretrained Vision-Language model (**VLM**) = vision encoder + text encoder fed to **AR-LLM**
2. Trained **diffusion**-based **Action Expert** : generates continuous agentic actions for robots

Hugging Face's SmolVLA is the open-source example:

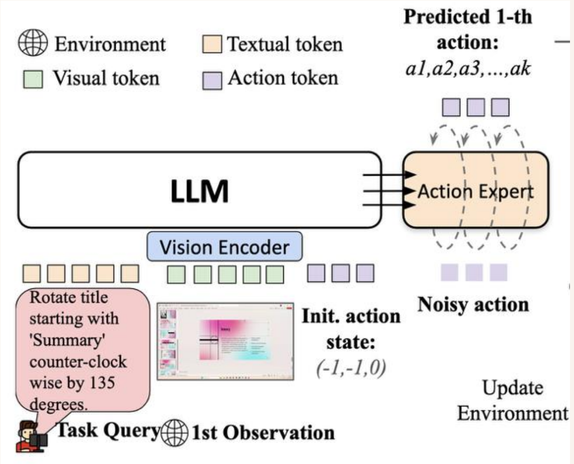


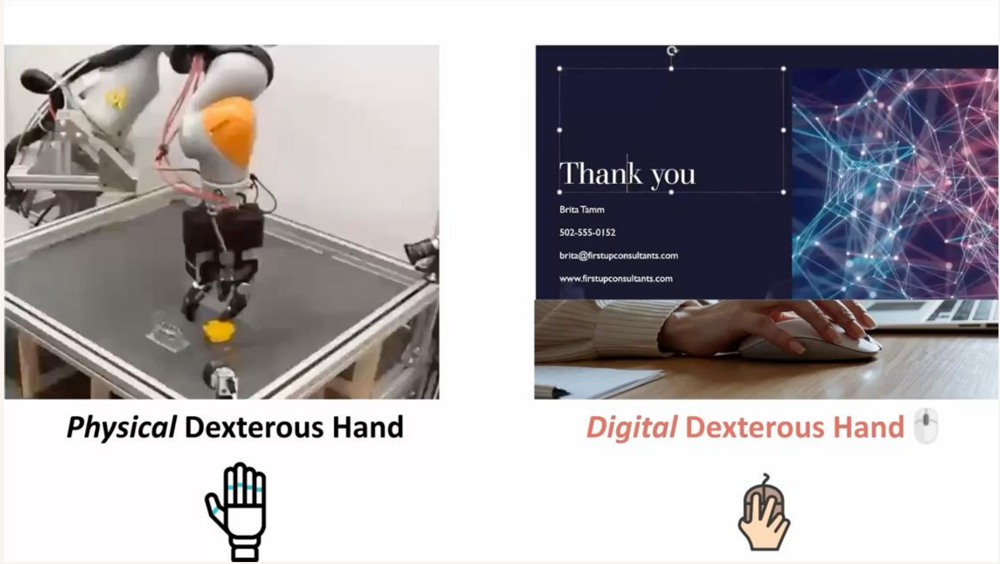
VLAs for Computer-use Agents

Multimodality of VLAs make them a good choice for **computer-use agents** (**CUA**) **w/o CLI**:

- Can we replace VLA's robotic embodiment with, e.g., **UI actions** in image-based agent?
- UI agents process the prompt and the UI **screenshot** to perform "clicks"

Example: ShowUI-Pi adopts the SmolVLA for the computer-use agent:



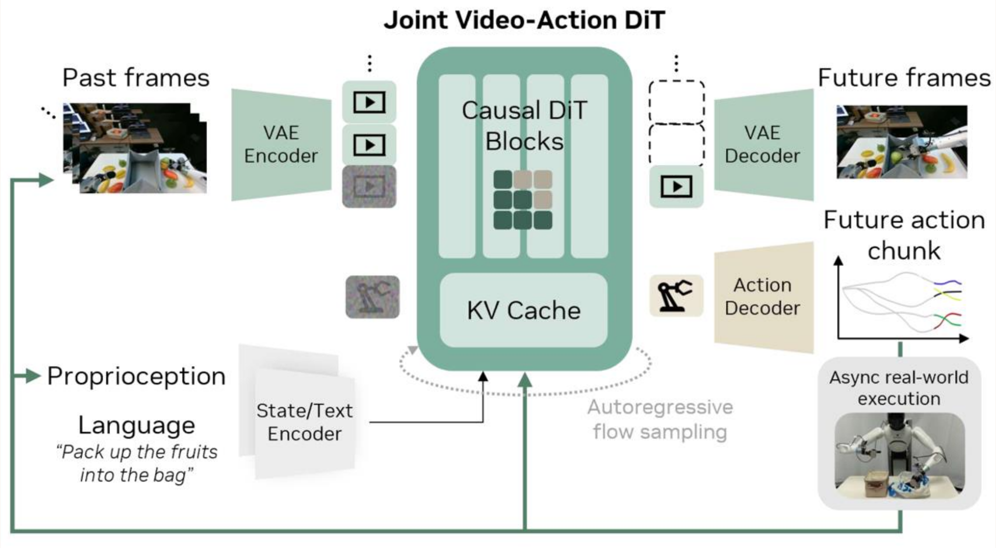


WAMs for Robotic Applications

Example: What is the next-generation World-Action model (**WAM**) ?

1. Pretrained image-to-video **diffusion backbone** predicts future actions and image frames
2. 2× improvement in **zero-shot generalization** to unseen prompts and actions w.r.t. VLAs

NVIDIA's recent DreamZero architecture:



Robust UI Agents using WAMs

We develop computer-use agent:

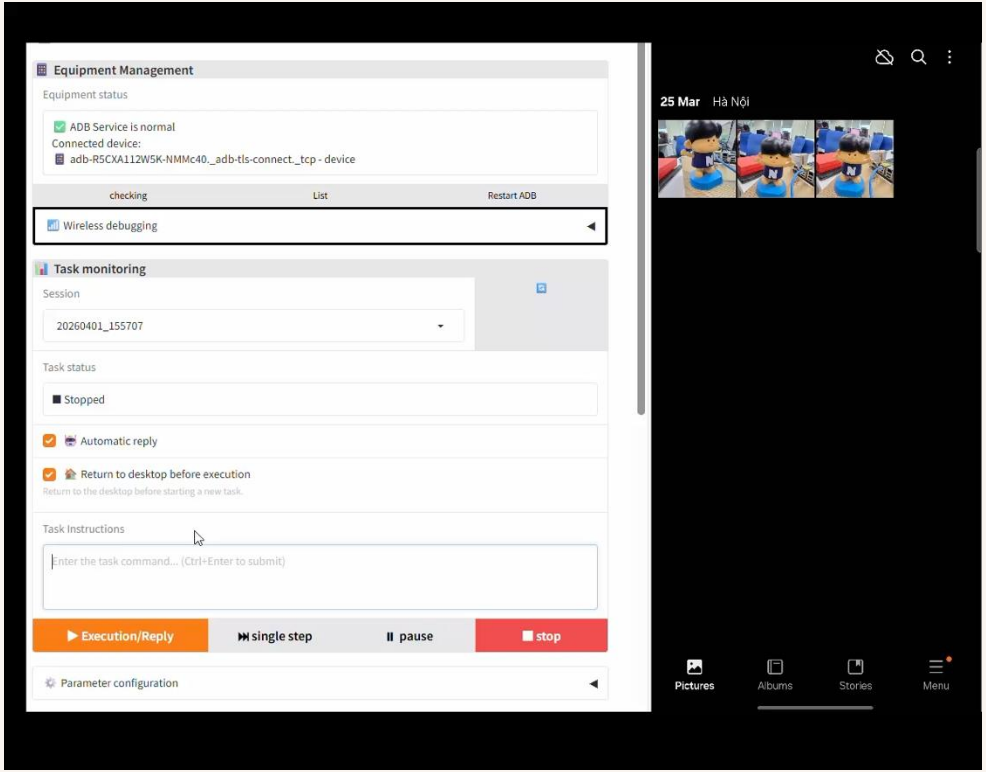
- Plenty of internal interfaces
- Often very **domain-specific**
- Require **100% robustness**
- **Low-complexity** on edge devices

Robust UI Agents using WAMs

We develop computer-use agent:

- Plenty of internal interfaces
- Often very **domain-specific**
- Require **100% robustness**
- **Low-complexity** on edge devices

Android Phone
@ ADB
Gallery
actions→



2. Technical

Accelerating Diffusion Models

Differential equations and few-step generation

Background

Diffusion models are trained to generate data from noise (or another data) via differential equation

$$\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$$

$\mathbf{x}_1 \sim p_{\text{data}}$

Forward process (fixed)

$\mathbf{x}_0 \sim \mathcal{N}(0, 1)$

Data

Noise

Reverse denoising process (generative)

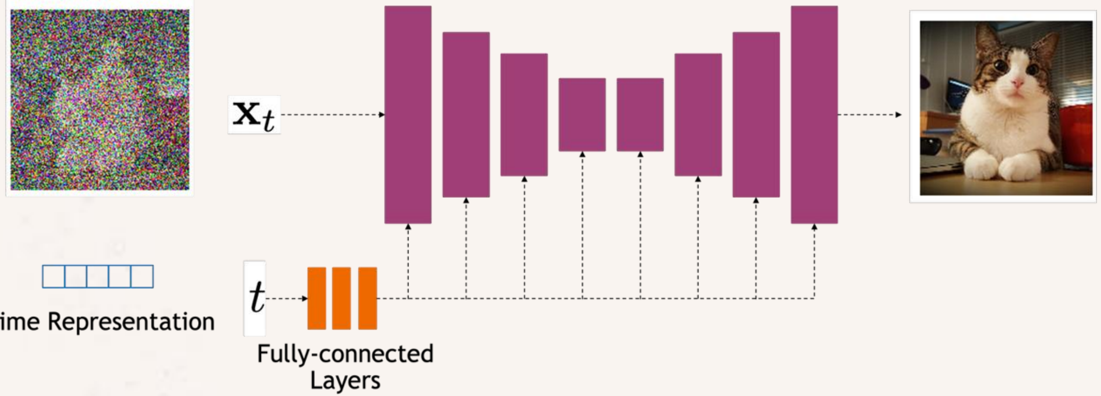
$$d\mathbf{x}_t = \mathbf{v}(\mathbf{x}_t, t)dt$$

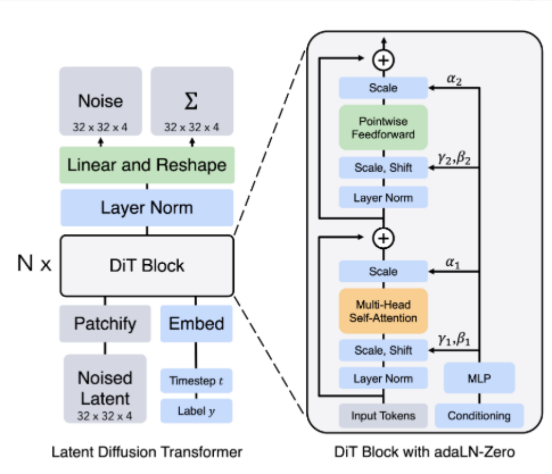
Lipman et al., Flow Matching for Generative Modeling, ICLR 2023
Albergo et al., Stochastic Interpolants: A Unifying Framework for Flows and Diffusions, ICLR 2024
Berner et al., Tutorial on Accelerated Diffusion Models, CVPR 2026

Architecture

Time-dependent architecture is used as the key difference from AR-LLMs

$$\boldsymbol{x}_1 = \int_{t=0}^{1} v_{\boldsymbol{\theta}}(t, \boldsymbol{x}_t)\,dt$$





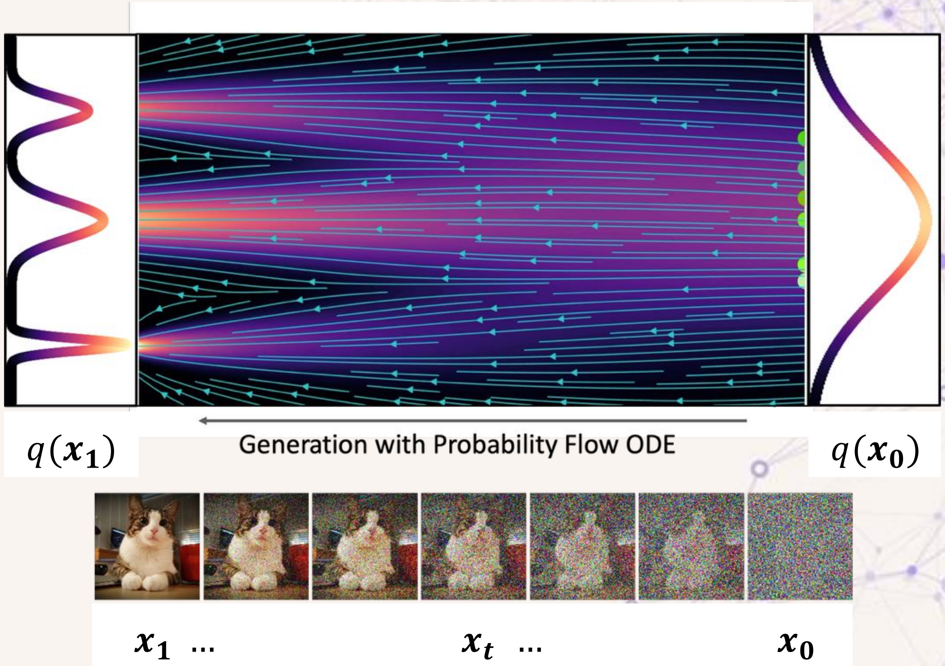
Peebles and Xie et al., Scalable Diffusion Models with Transformers, ICCV 2023
Berner et al., Tutorial on Accelerated Diffusion Models, CVPR 2026

Few-step Generation

$$x_1 = \int_{t=0}^{1} v_{\theta}(t, x_t)dt$$

Integration is computational challenging

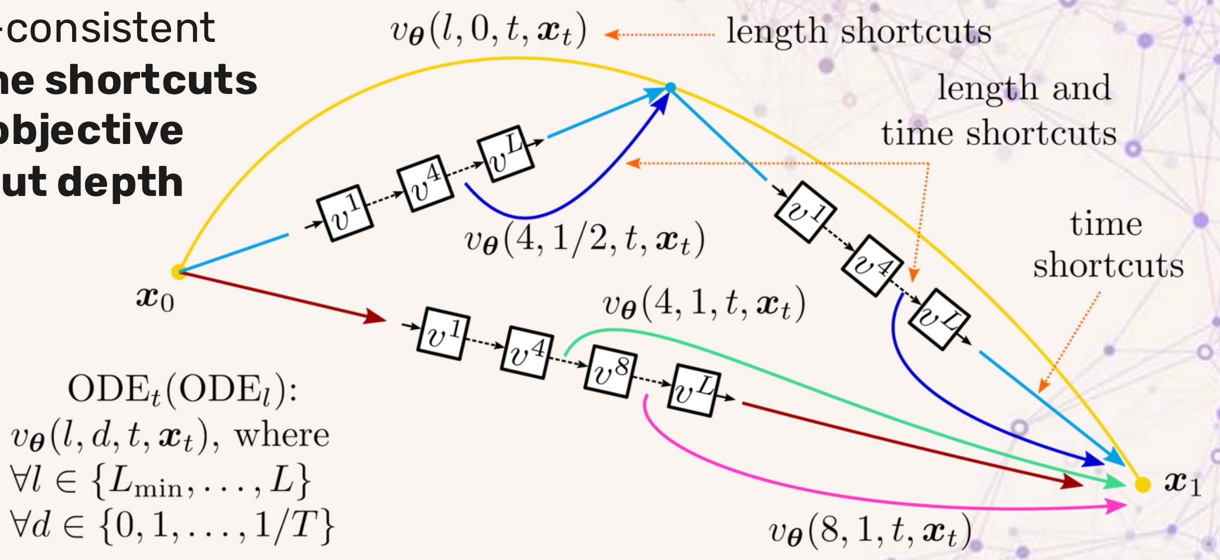
1. Time **discretization** reduces complexity
2. Number of **time steps** can be minimized
3. Eventually, one- or few-**steps** is enough



Frans et al., One Step Diffusion via Shortcut Models, ICLR 2025
Zhou et al., Inductive Moment Matching, ICML 2025
Geng et al., Mean Flows for One-step Generative Modeling, NeurIPS 2025

Time and Depth Shortcuts

- Discretized trajectories are self-consistent with the full trajectory using **time shortcuts**
- Training imposes **consistency objective**
- In our approach, we also **shortcut depth or length** of the neural network

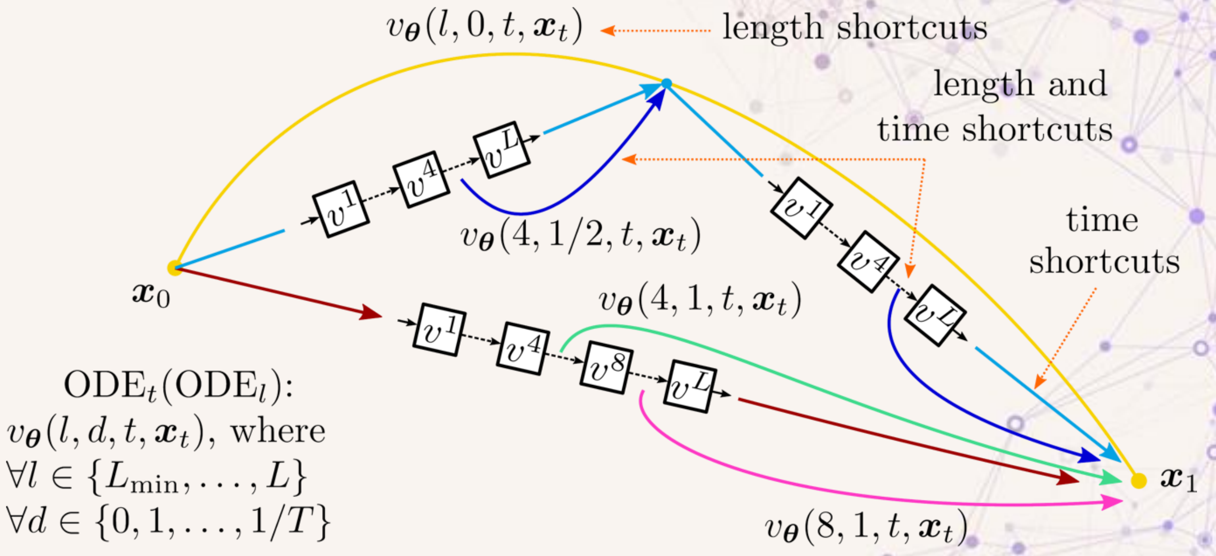


Frans et al., One Step Diffusion via Shortcut Models, ICLR 2025
Zhou et al., Inductive Moment Matching, ICML 2025
Geng et al., Mean Flows for One-step Generative Modeling, NeurIPS 2025
Gudovskiy et al., Shortcutting the Time and the Length in Diffusion and Flow Models, WACV 2026

Time and Depth Shortcuts

Results:

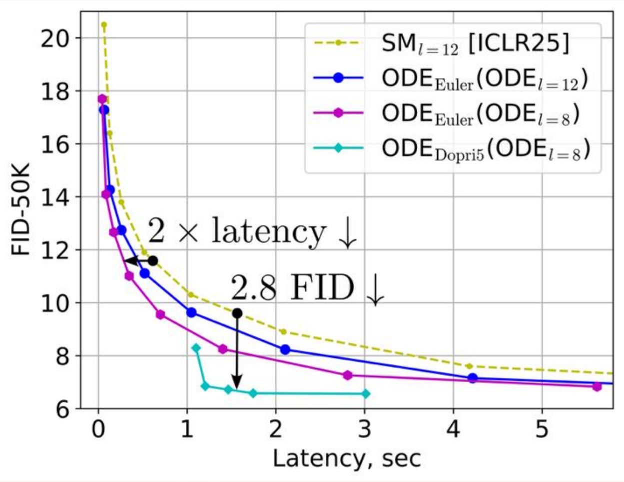
- **Complexity-quality** control
- 10-100x lower **latency**
- 1.5-2x less **memory size** in the depth shortcut approach

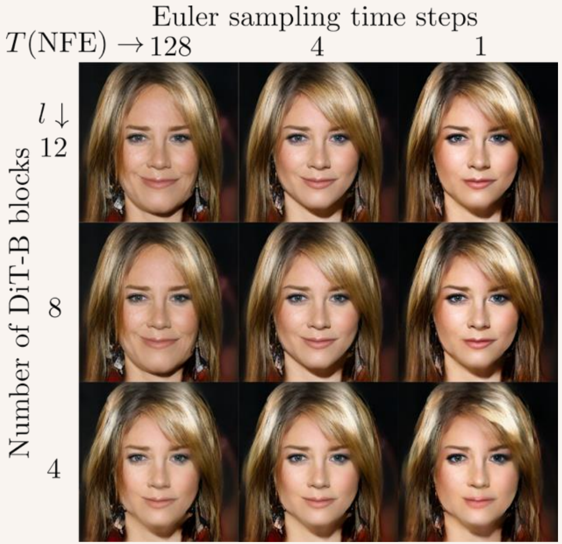


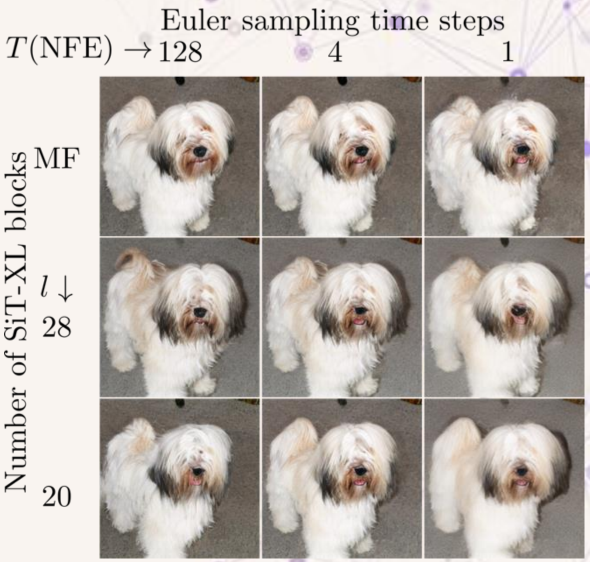
Frans et al., One Step Diffusion via Shortcut Models, ICLR 2025
Zhou et al., Inductive Moment Matching, ICML 2025
Geng et al., Mean Flows for One-step Generative Modeling, NeurIPS 2025
Gudovskiy et al., Shortcutting the Time and the Length in Diffusion and Flow Models, WACV 2026

Time and Depth Shortcuts

Time steps T (NFE) vs. network depth l







Gudovskiy et al., Shortcutting the Time and the Length in Diffusion and Flow Models, WACV 2026

3. Outro
Conclusions

Next-generation agents and diffusion models

Next-Gen Agentic Models

AR-LLMs, VLMs

- ✓ Continue to scale in CLI-type agentic apps
- ❖ Struggle with "continuous" data
- ❖ Less flexible in selecting complexity vs. quality
- ✓ Hybrid VLAs extend agentic data domains to robots and CUAs

dLLMs, WAMs

- ✓ dLLMs are fast and flexible in token generation, e.g., blocks
- ❖ dLLMs might be limited in reasoning abilities
- ✓ WAMs excel in zero-shot generalization
- ❖ WAMs are currently less optimized than VLAs

Multi- / Omnimodal

- ➢ Heterogeneous data push for hybrid models
- ➢ Multi- and omnimodal agents combine the best architectures
- ➢ Domain-specific agents increase robustness with a fraction of the "generalist" cost

Reducing Costs in Agentic Apps

AR vs. DM

- ✓ Compared to AR-LLMs, DM's approach is generally more efficient and outperforms in many data domains.
- ❖ However, even **DM complexity** is a challenge for real-world deployments

Few-step sampling

- ✓ Recent few-step sampling methods show up to **100x** reduction in latency with minor degradation
- ❖ But the **memory bottleneck** is not addressed with fewer time steps

Our depth shortcuts

- ✓ Our approach generalizes few-step sampling to the network depth dimension
- ✓ This allows to shrink memory size by up to **2x** which is crucial for real-world edge CUAs with the quality control

Ai4

Denis Gudovskiy
Distinguished AI Engineer

Questions?

Denis Gudovskiy
denis.gudovskiy@us.panasonic.com