DDiT: Dynamic Patch Scheduling for Efficient Diffusion Transformers

CVPR 2026 Highlight

Dahye Kim[1,2] Deepti Ghadiyaram[2] Raghudeep Gadde[1]

[1]Amazon [2]Boston University

{dahye, dghadiya}@bu.edu raghudeep.g@gmail.com

Inference in Diffusion Models is Expensive 💰

- Generating a 5-second 720p video with Wan-2.1 takes ~**30 minutes***.

*On RTX 4090

Inference in Diffusion Models is Expensive 💰

- Generating a 5-second 720p video with Wan-2.1 takes ~**30 minutes***.
- Reasons:
  - Diffusion models use an iterative denoising process
  - Each step requires a full forward pass over all spatial tokens

Noisy image | Less noisy image | Even less noisy image | Generated image

*On RTX 4090

1. *Should every image be encoded at the same granularity?*

Images have very different semantic complexity

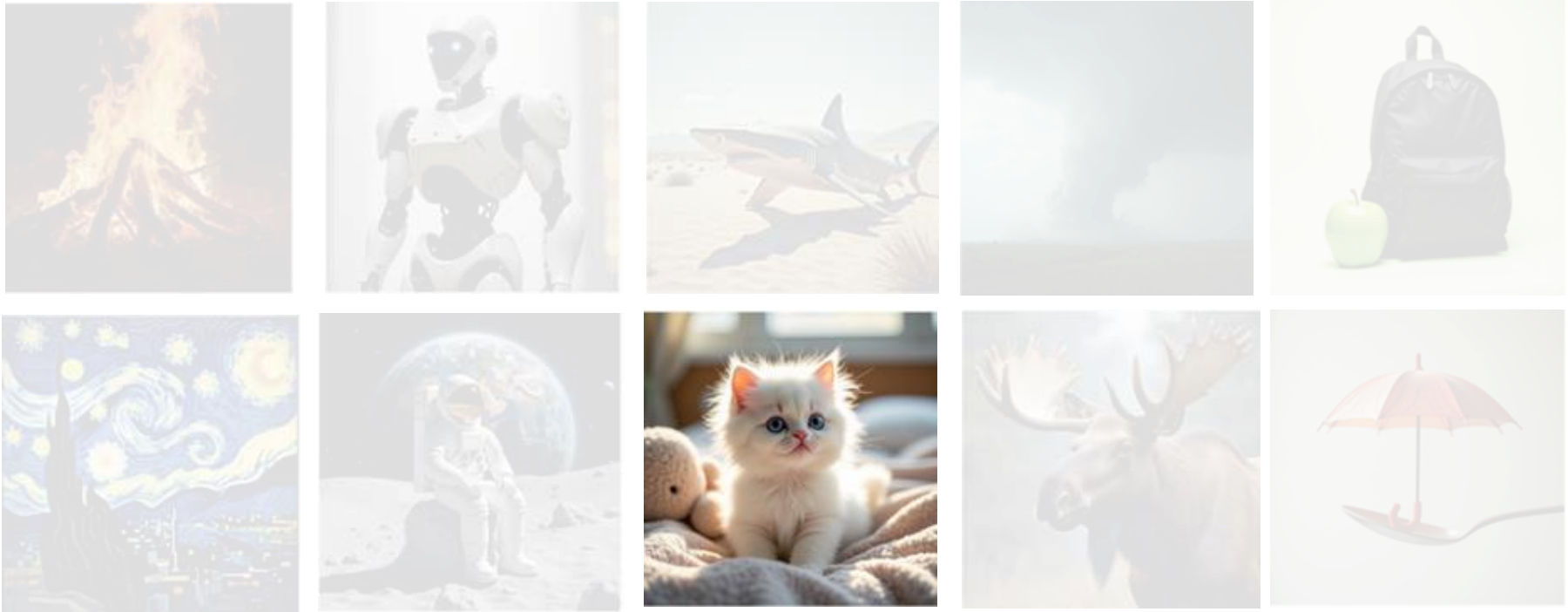






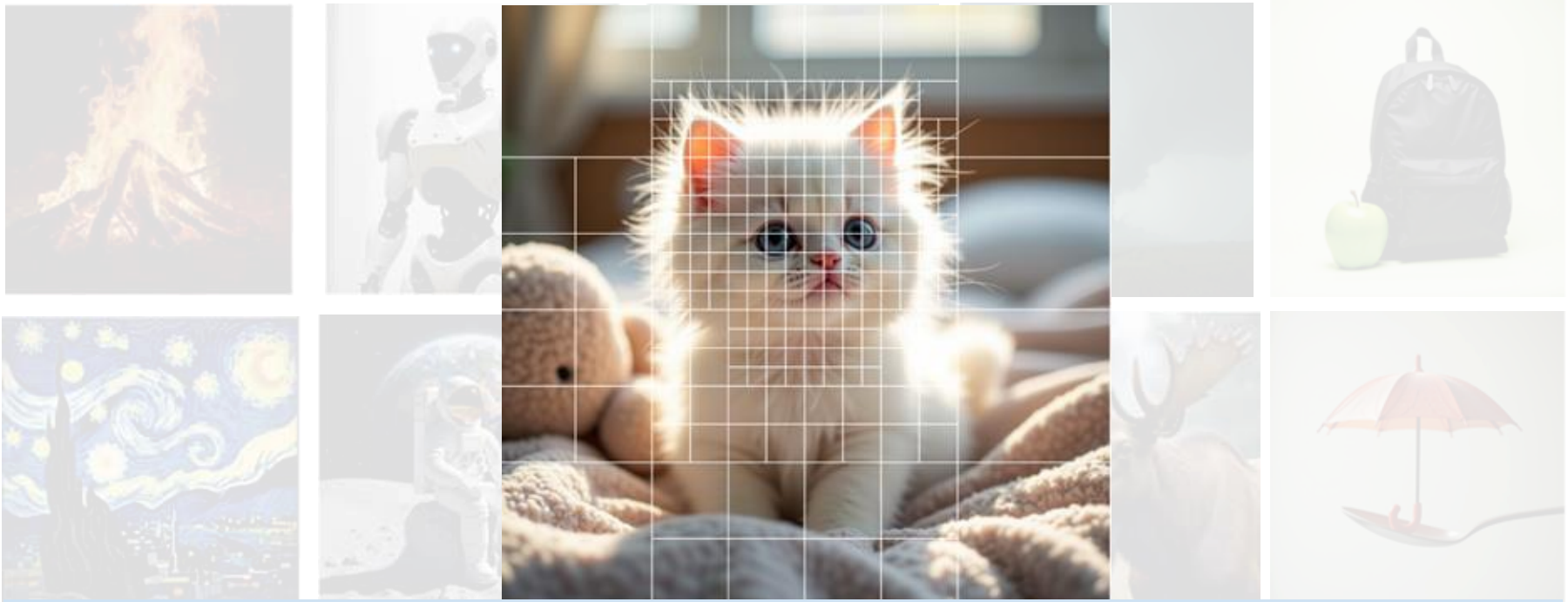



Images have very different semantic complexity



Images have very different semantic complexity

Images have very different semantic complexity

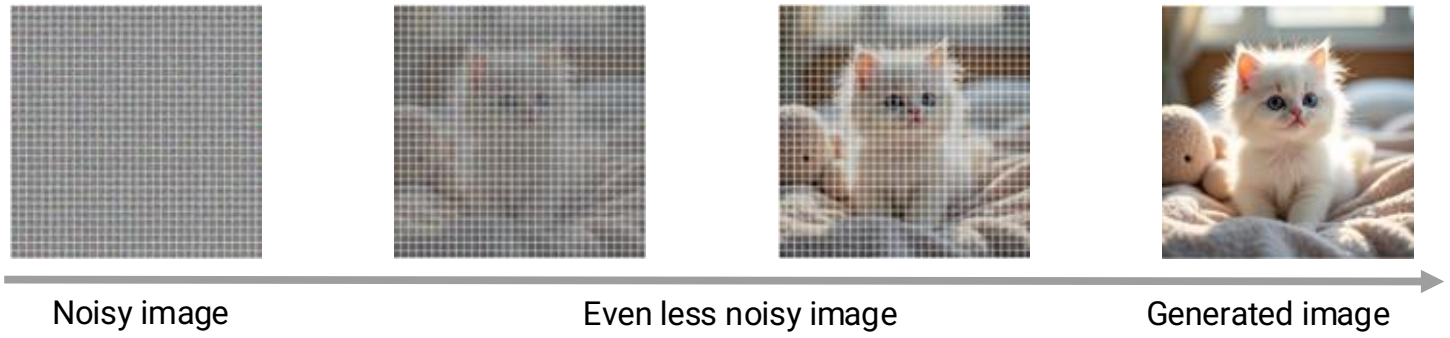


Takeaway: Different regions in the same image can be encoded at different spatial granularity.

1. *Should every image be encoded at the same granularity?*

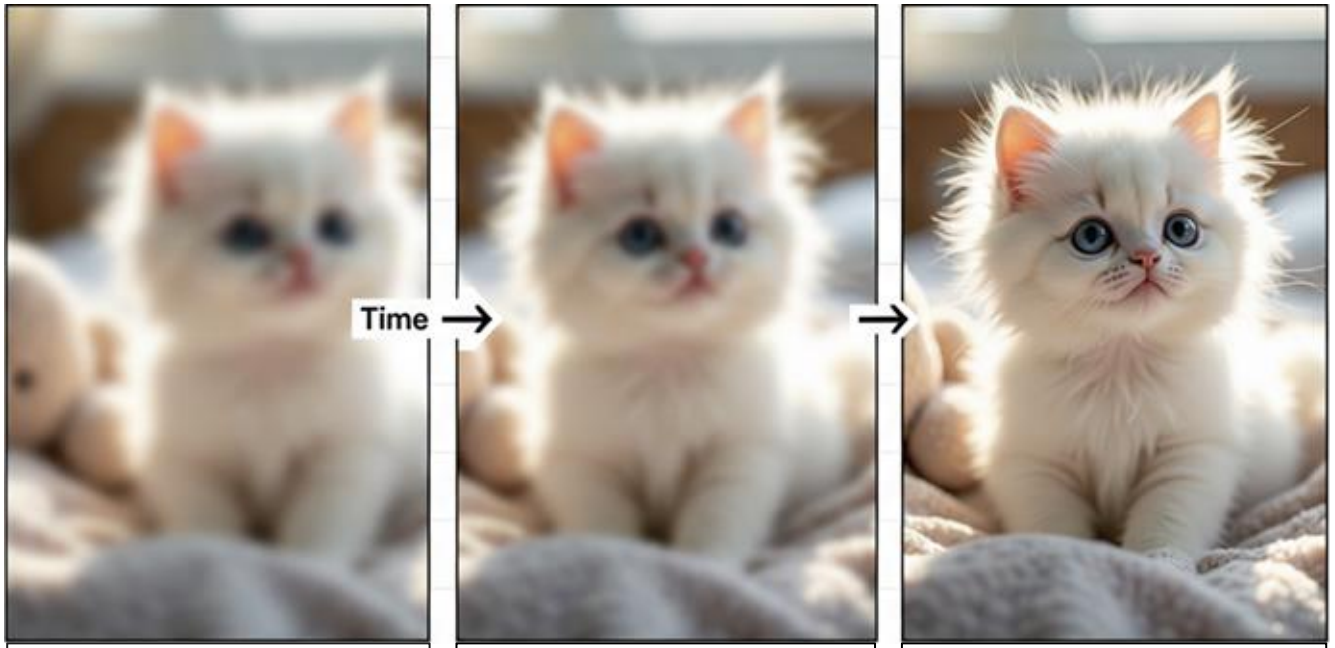
Not necessarily…

2. Should every denoising step process the image at the same granularity?



2. Should every denoising step process the image at the same granularity?

Denoising



Coarse:
Layout & composition

Intermediate:
Object geometry

Fine-grained:
Sharpened textures

[1] D. Kim et al., Revelio: Interpreting and leveraging semantic information in diffusion models. ICCV'2025
[2] S. Mahajan et al., Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. CVPR'2024
[3] O. Patashnik et al., Localizing object-level shape variations with text-to-image diffusion models. ICCV'2023
[4] L. Tang et al., Emergent correspondence 931 from image diffusion. NeurIPS'2023

2. Should every denoising step process the image at the same granularity?

Not necessarily…

Inference in Generative Models is expensive

1. *Should every image be encoded at the same granularity?*

Not necessarily…

2. *Should every denoising step process the image at the same granularity?*

Inference in Generative Models is expensive

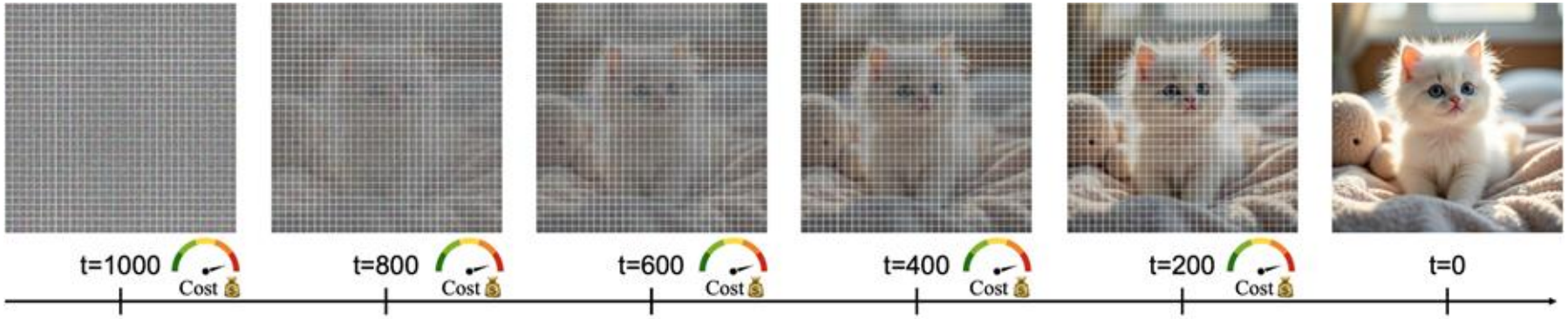
1. *Should every image be encoded at the same granularity?*

2. Should every denoising step process the image at the same granularity?

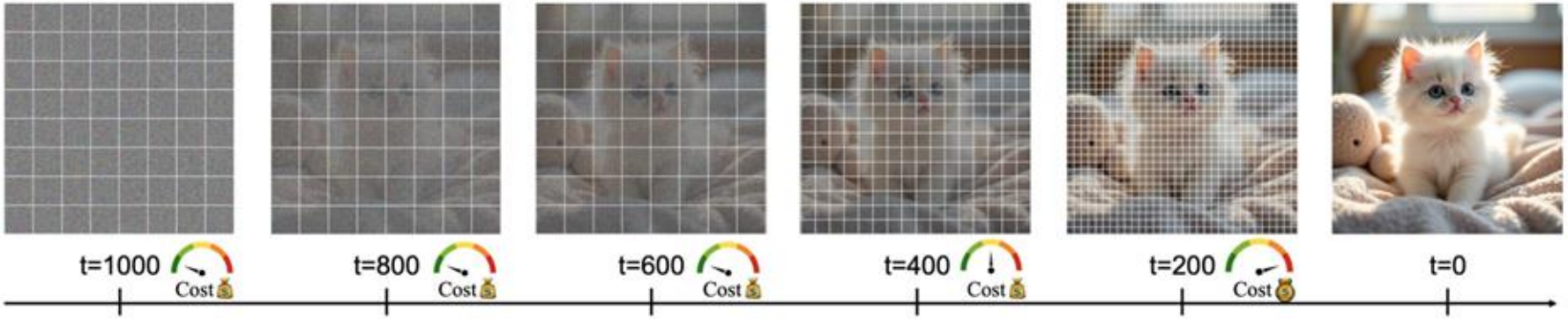
*Key idea: Dynamically vary patch sizes based on **content complexity** and the denoising timestep.*

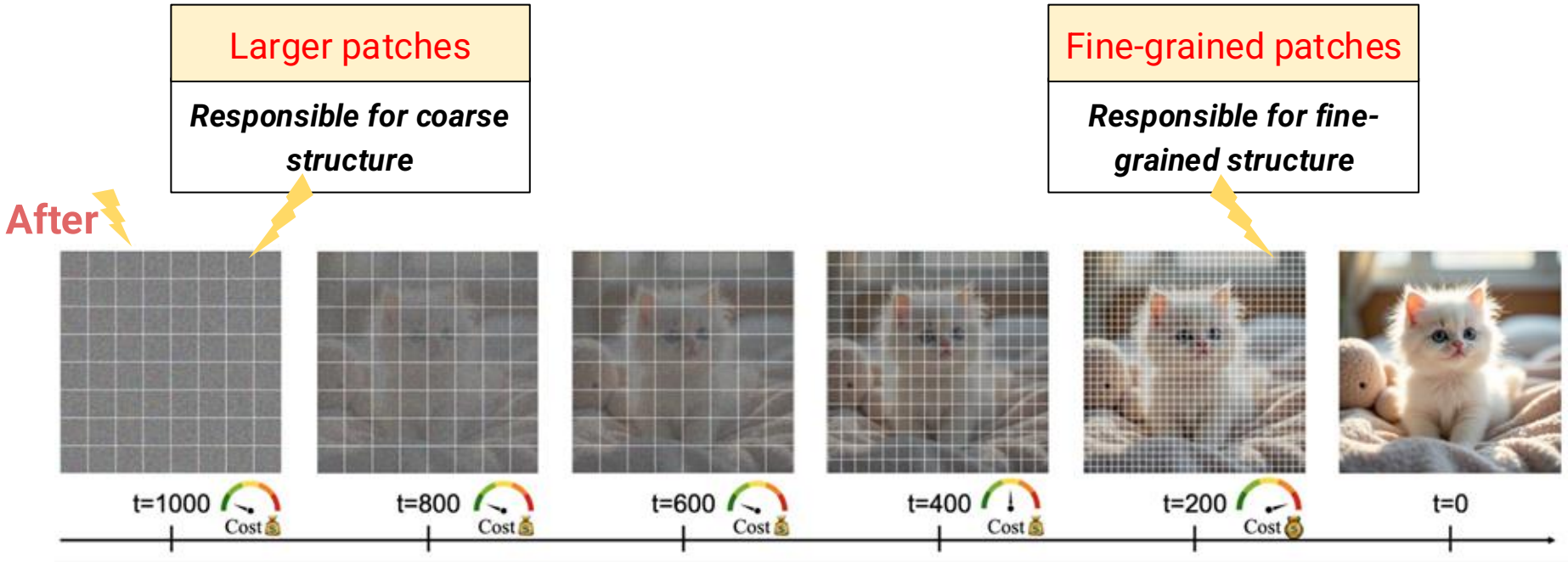
***Dynamically** vary patch sizes based on **content** and **timestep**.*

Baseline



After




Larger patches
Responsible for coarse structure
Fine-grained patches
Responsible for fine-grained structure
After
t=1000
Cost
t=800
Cost
t=600
Cost
t=400
Cost
t=200
Cost
t=0

Can we use latent's rate of change as a proxy for patch size?

How to determine the optimal patch size at each timestep?

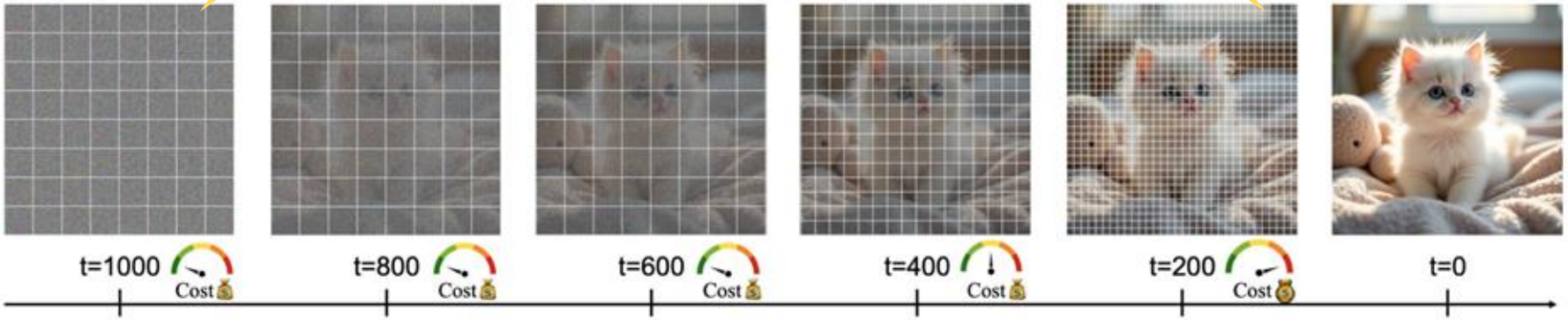
Larger patches

Responsible for coarse structure

Fine-grained patches

Responsible for fine-grained structure

After



How to measure a latent's evolution?

t-1 t t+1

z_{t-1} z_t z_{t+1}

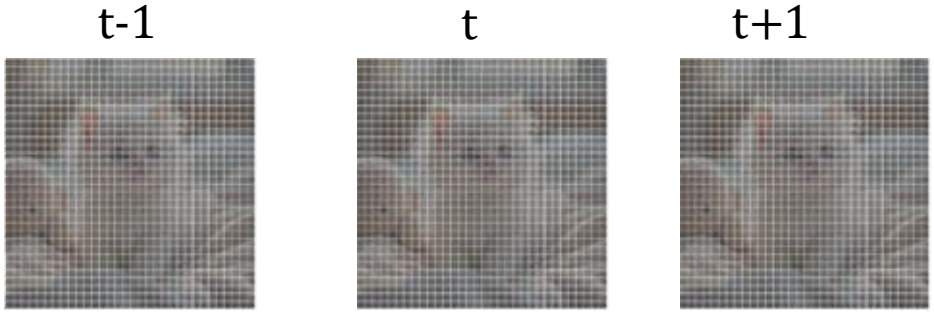
Displacement $\Delta \mathbf{z}_t = \mathbf{z}_t - \mathbf{z}_{t+1}.$

Velocity $\Delta^{(2)} \mathbf{z}_{t-1} = \Delta \mathbf{z}_{t-1} - \Delta \mathbf{z}_t.$

Acceleration $\Delta^{(3)} \mathbf{z}_{t-1} = \Delta^{(2)} \mathbf{z}_{t-1} - \Delta^{(2)} \mathbf{z}_t$ Rate of change of latent.

How to measure a latent's evolution?

Patch size = 2



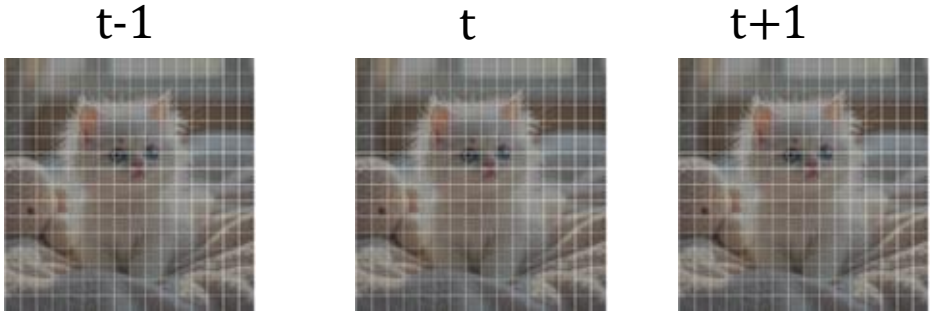
Acceleration $\Delta^{(3)}\mathbf{z}_{t-1} = \Delta^{(2)}\mathbf{z}_{t-1} - \Delta^{(2)}\mathbf{z}_t$

Compute standard deviation of acceleration at different patch sizes

Low standard deviation? Latent evolves slowly -> go more coarse grained.

How to measure a latent's evolution?

Patch size = 4



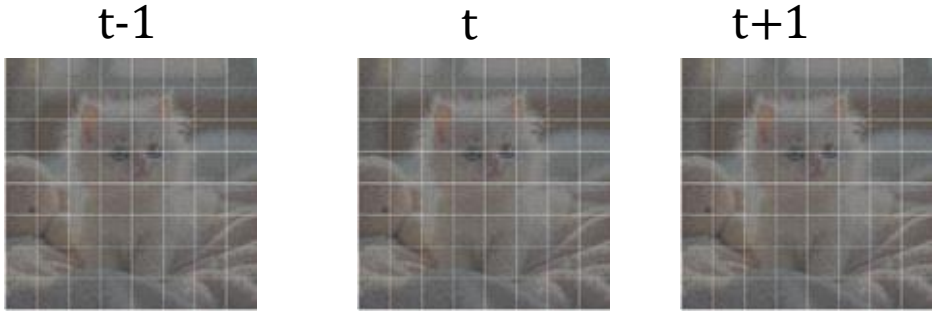
Acceleration $\Delta^{(3)}\mathbf{z}_{t-1} = \Delta^{(2)}\mathbf{z}_{t-1} - \Delta^{(2)}\mathbf{z}_t$

Compute standard deviation of acceleration at different patch sizes

Low standard deviation? Latent evolves slowly -> go more coarse grained.

How to measure a latent's evolution?

Patch size = 8

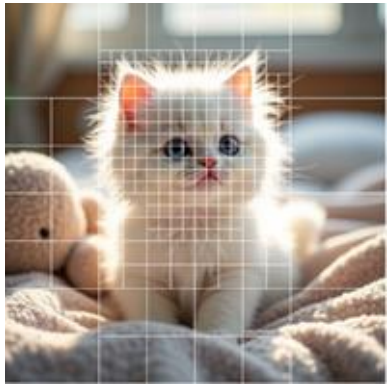


Acceleration $\Delta^{(3)}\mathbf{z}_{t-1} = \Delta^{(2)}\mathbf{z}_{t-1} - \Delta^{(2)}\mathbf{z}_t$

Compute standard deviation of acceleration at different patch sizes

How to measure a latent’s evolution?

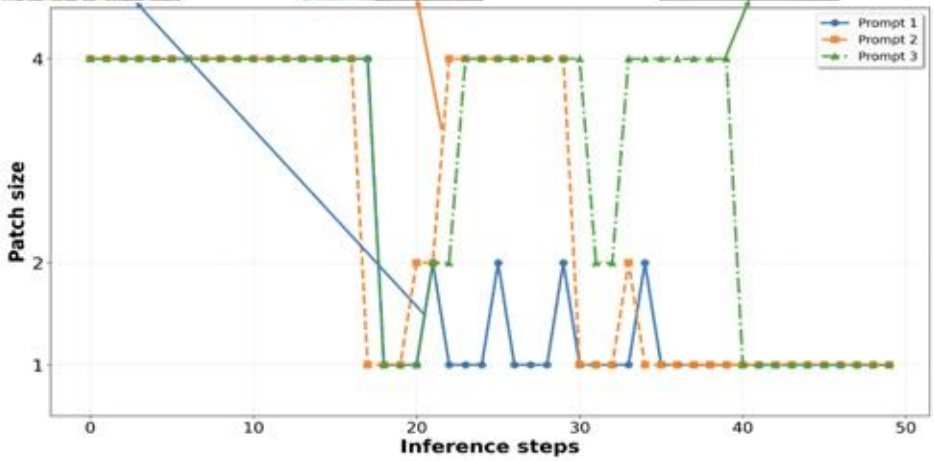
- Hypothesis: Fine-grained image region leads to rapid latent evolution



Solution: **Dynamic patch scheduler**

1. Use standard deviation of acceleration at different patch sizes.
2. Determine the right patch size.

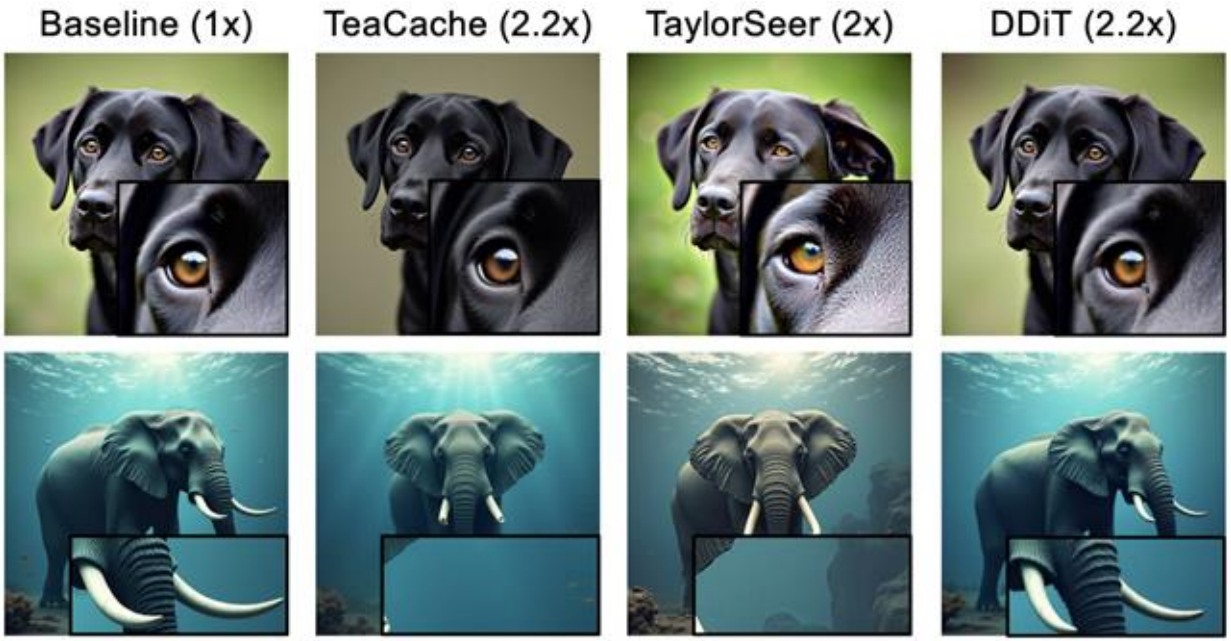
adapts to each



Observations:

1. High textured regions take smaller patches ⇒ same compute as baseline.
2. Smooth regions take larger patches ⇒ less compute

Results: Text to Image Generation



DDiT preserves

- ✅ fine-grained details
- ✅ pose
- ✅ spatial layout, and
- ✅ overall color distribution

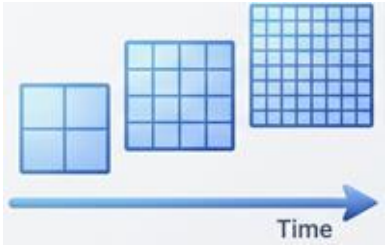
[5] F. Liu et al., Timestep embedding tells: It's time to cache for video diffusion model. CVPR'2025
[6] J. Liu et al., From reusing to forecasting: Accelerating diffusion models with taylorseers. ICCV'2025
[7] Black Forest Labs. FLUX. 2024

Results: Text to Video Generation

✅ 3.2X speedup
✅ Maintains perceptual quality

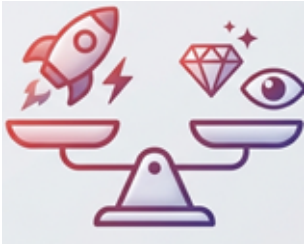
[8] T. Wan et al., Wan: Open and advanced large-scale video generative models. arXiv'2025

Summary: DDiT



Adaptive Granularity

Dynamic Patching



Max Efficiency & Quality