ReFrame:
Layer Caching for Accelerated Inference in Real-Time Rendering

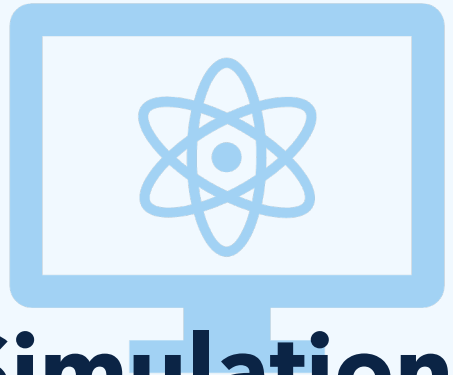
Lufei Liu and Tor M. Aamodt

The University of British Columbia

Real-Time Rendering

AR/VR

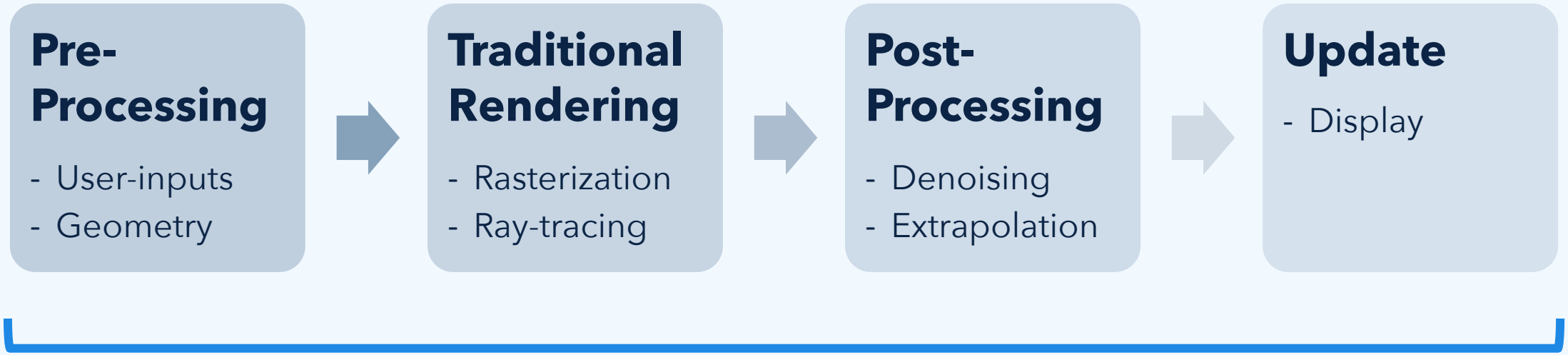
Gaming



Simulations

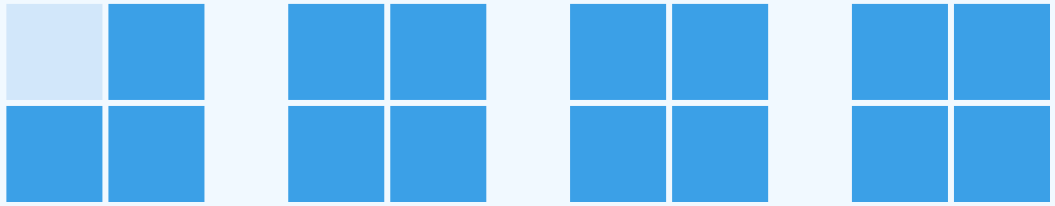
3D Design

Real-Time Rendering



< 16.7 ms $\approx$ 60 FPS

< 11.1 ms $\approx$ 90 FPS

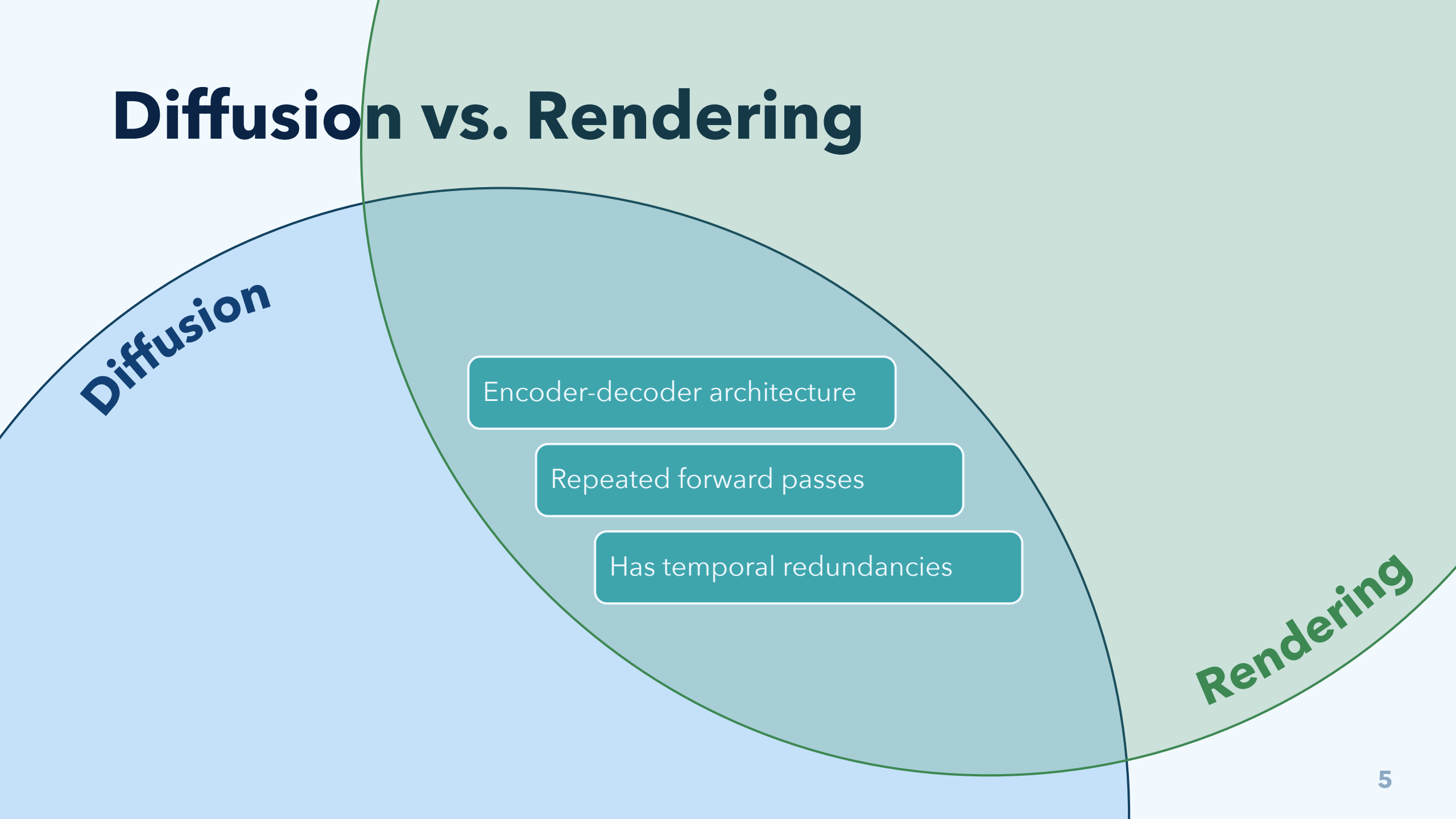


15 out of every 16 pixels are AI-generated

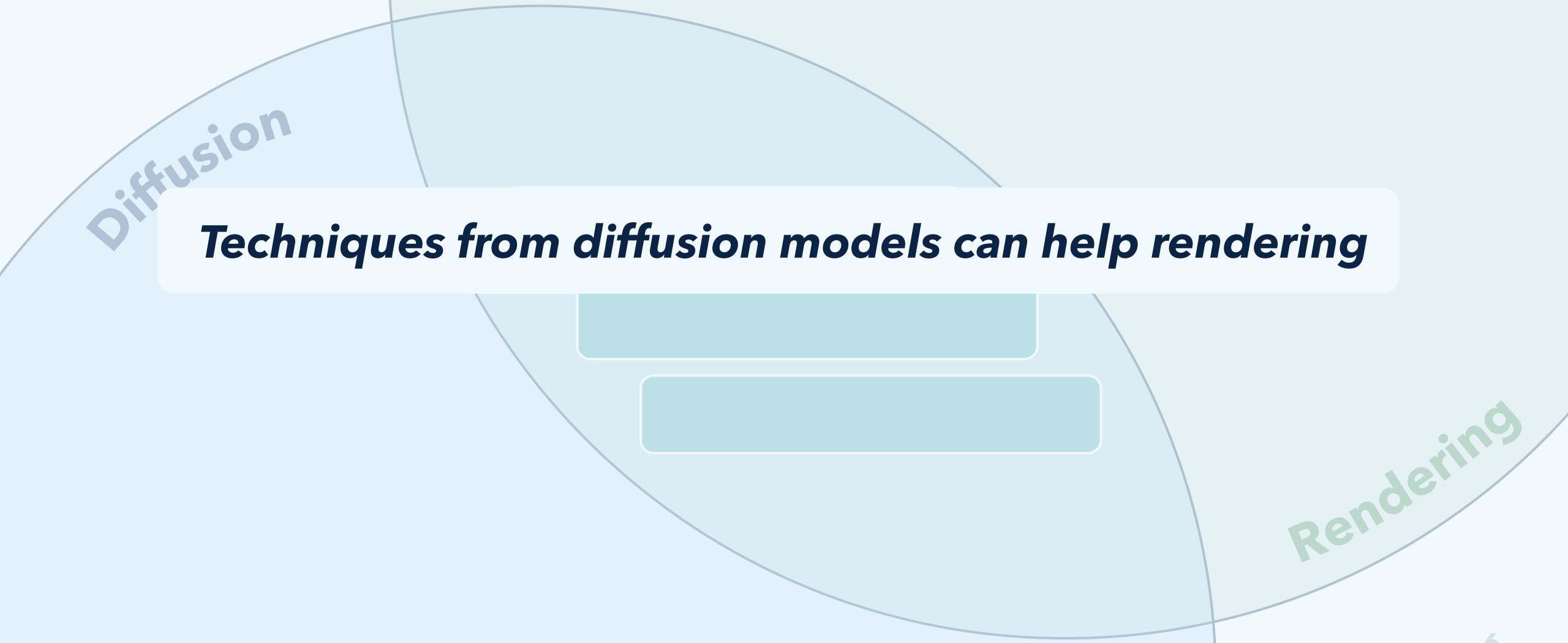
Accelerating neural network inferences is important for better graphics

DLSS 4.0: https://www.nvidia.com/en-us/geforce/news/dlss4-multi-frame-generation-ai-innovations/

Diffusion vs. Rendering



Diffusion vs. Rendering



Techniques from diffusion models can help rendering

Diffusion vs. Rendering

Diffusion

Predetermined forward passes

Only final inference counts

Image quality > inference time

Encoder-decoder architecture

Repeated forward passes

Has temporal redundancies

Rendering

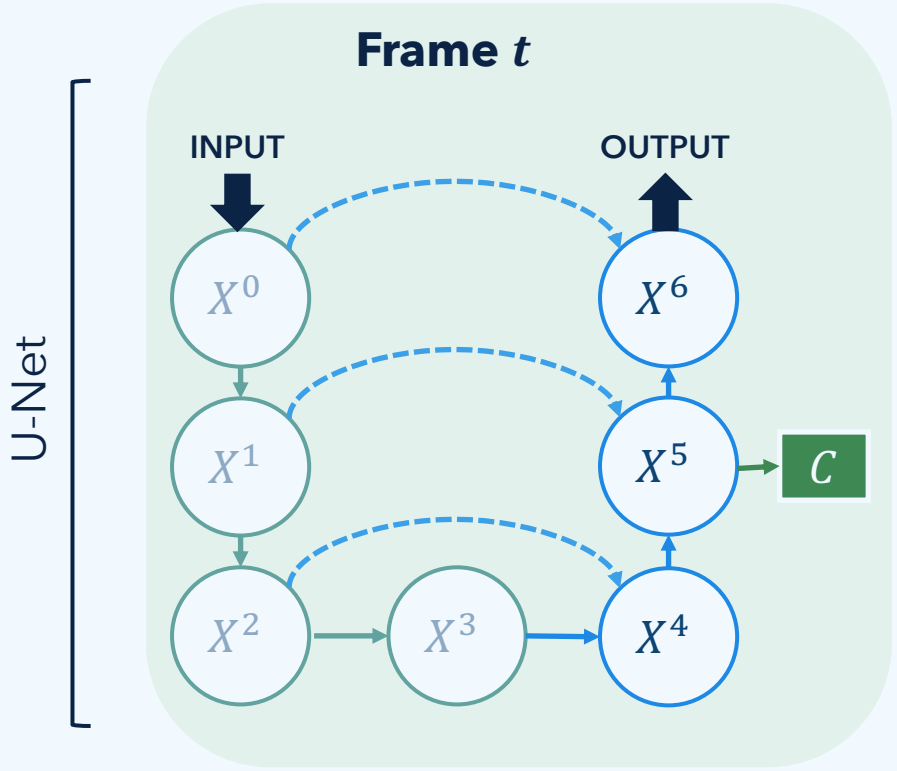
Random behavior from user

Every inference matters

Strict inference time enforced

ReFrame

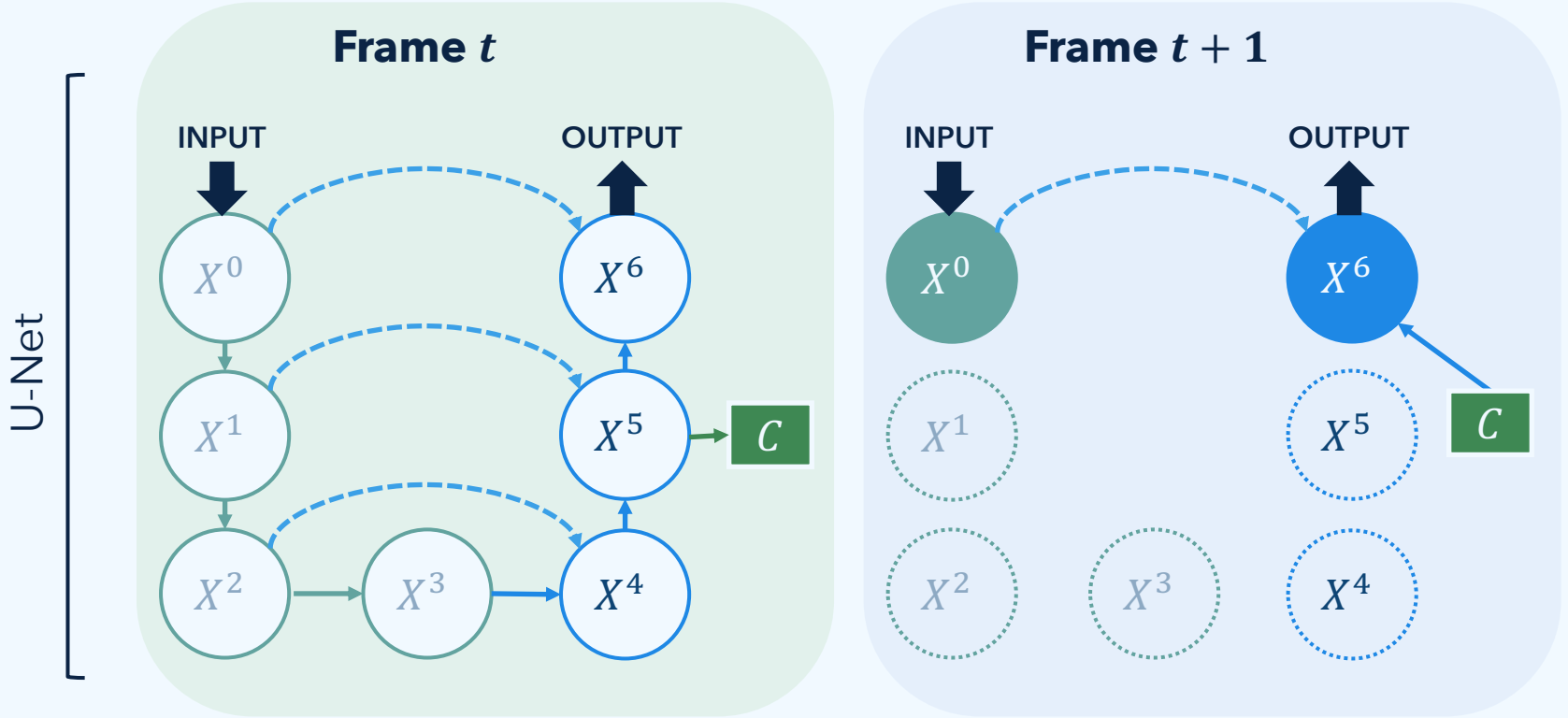
We extend the caching scheme introduced by **DeepCache** [1] to rendering workloads



*[1] Xinyin Ma, Gongfan Fang, and Xinchao Wang. "**DeepCache: Accelerating Diffusion Models for Free.**" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.*

ReFrame

We extend the caching scheme introduced by **DeepCache** [1] to rendering workloads

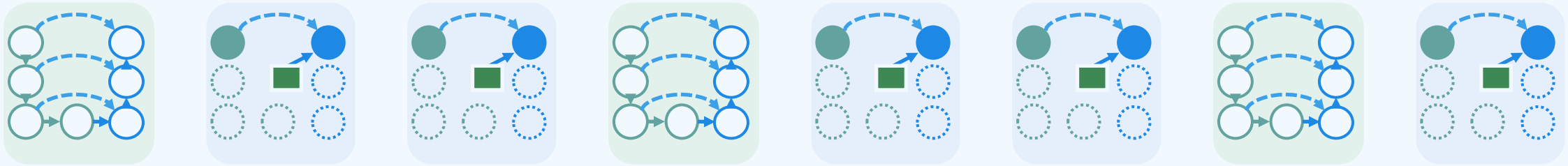


*[1] Xinyin Ma, Gongfan Fang, and Xinchao Wang. "**DeepCache: Accelerating Diffusion Models for Free.**" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.*

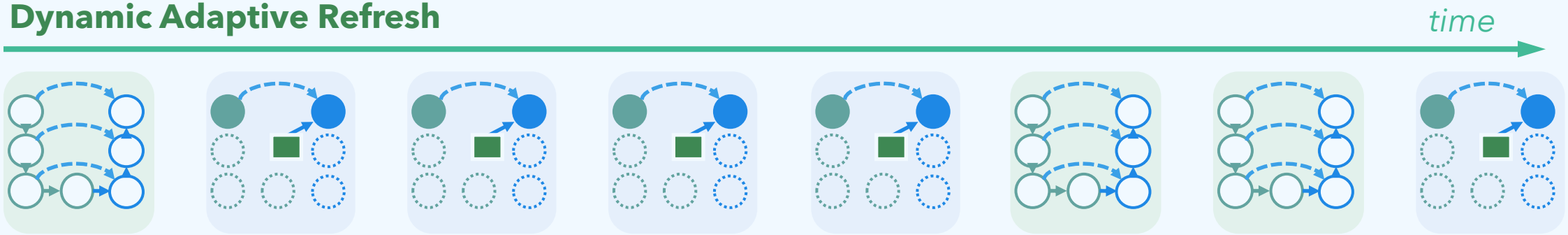
ReFrame

When should the cache be refreshed?

✖ **Static Every-N Refresh**



✔ **Dynamic Adaptive Refresh**

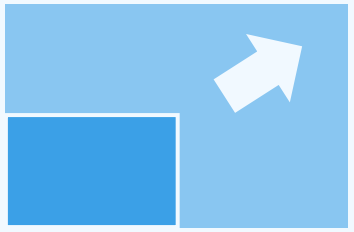


Evaluation

End-to-end network inference executed on NVIDIA RTX 2080 GPU

Frame Extrapolation

ExtraNet
SIGGRAPH 2021
[Guo et al.]



Supersampling

Fourier-Based Super Resolution
SIGGRAPH 2024
[Zhang et al.]

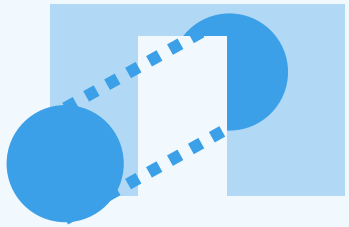


Image Composition

Implicit Depth
CVPR 2023
[Watson et al.]

ReFrame achieves:

Up to 1.05–1.85x inference speedup

***with* negligible FLIP error of 0.006–0.1**

***by* extending existing caching techniques with dynamic policies**

Visit https://ubc-aamodt-group.github.io/reframe-layer-caching/ for more details!