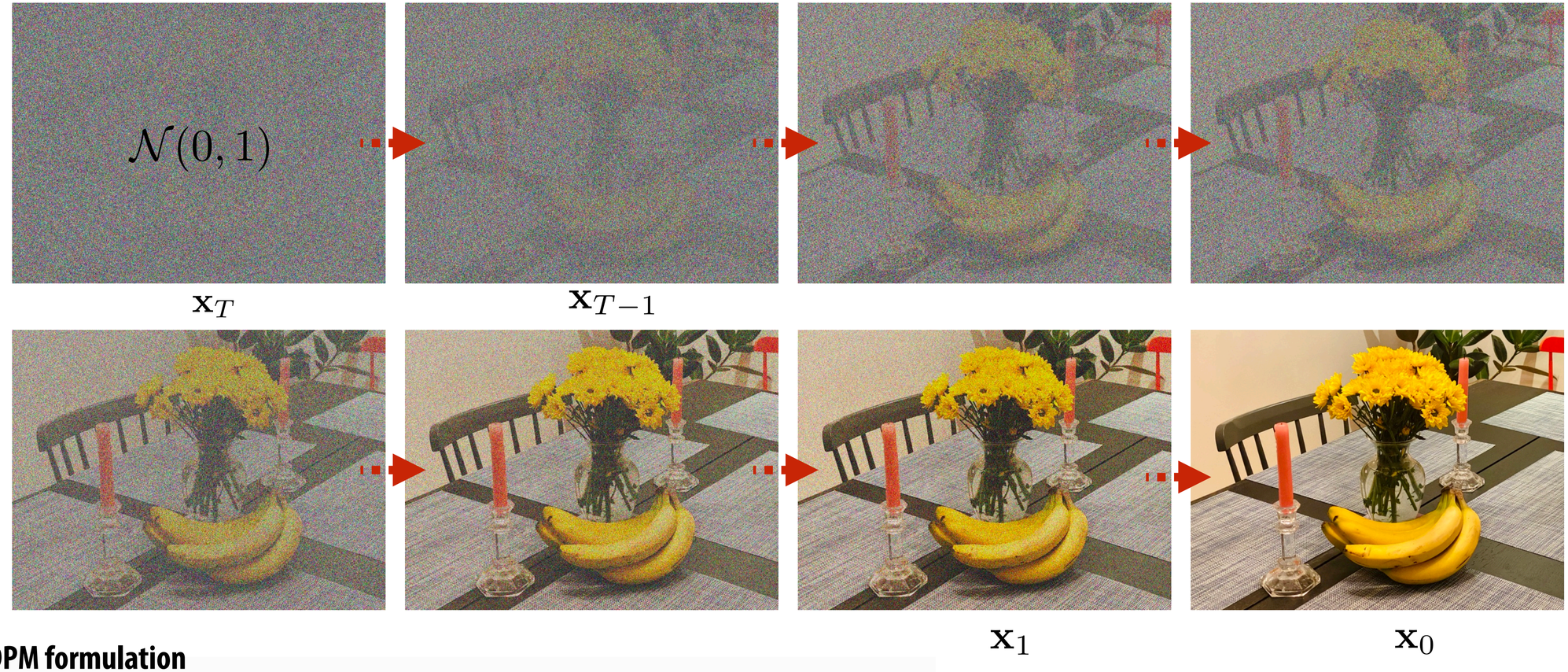
Lecture 18:

Performance Optimization of Autoregressive Video Models + Course Recap

Visual Computing Systems
Stanford CS348K, Spring 2026

Diffusion review

- Iterative process (T steps) to transform noise x_T into a sample X_0 from a given data distribution
- Denoising function given by a learned neural network



Common DDPM formulation

```
p_θ(x_{t-1} | x_t) = N(x_{t-1};  μ_θ(x_t, t),  Σ_θ(x_t, t))
```

Transformer review

- Project sequence of N input tokens into Q, K, V
- Compute attention

Let N be the length of the input token sequence

Let Q be a N x d matrix — "queries"
Let K be a N x d matrix — "keys"
Let V be a N x d matrix — "values"
(Q,K,V) = project(tokens)

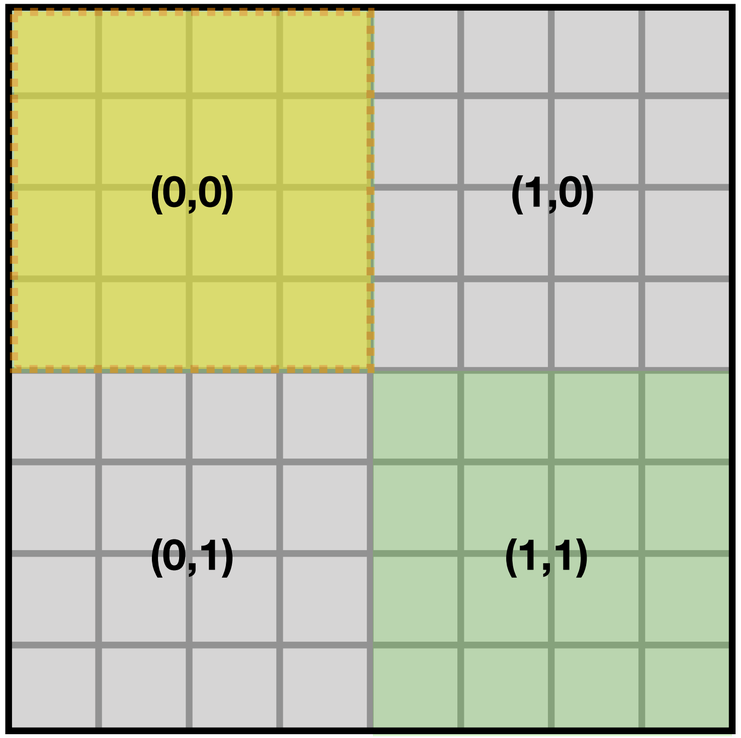
Let $\mathbf{S} = \mathbf{Q}\mathbf{K}^\top \in \mathbb{R}^{N \times N}$

Let $\mathbf{P} = \mathrm{softmax}(\mathbf{S}) \in \mathbb{R}^{N \times N}$ $\mathrm{softmax}(\mathbf{S})$ is computing softmax over the rows of S

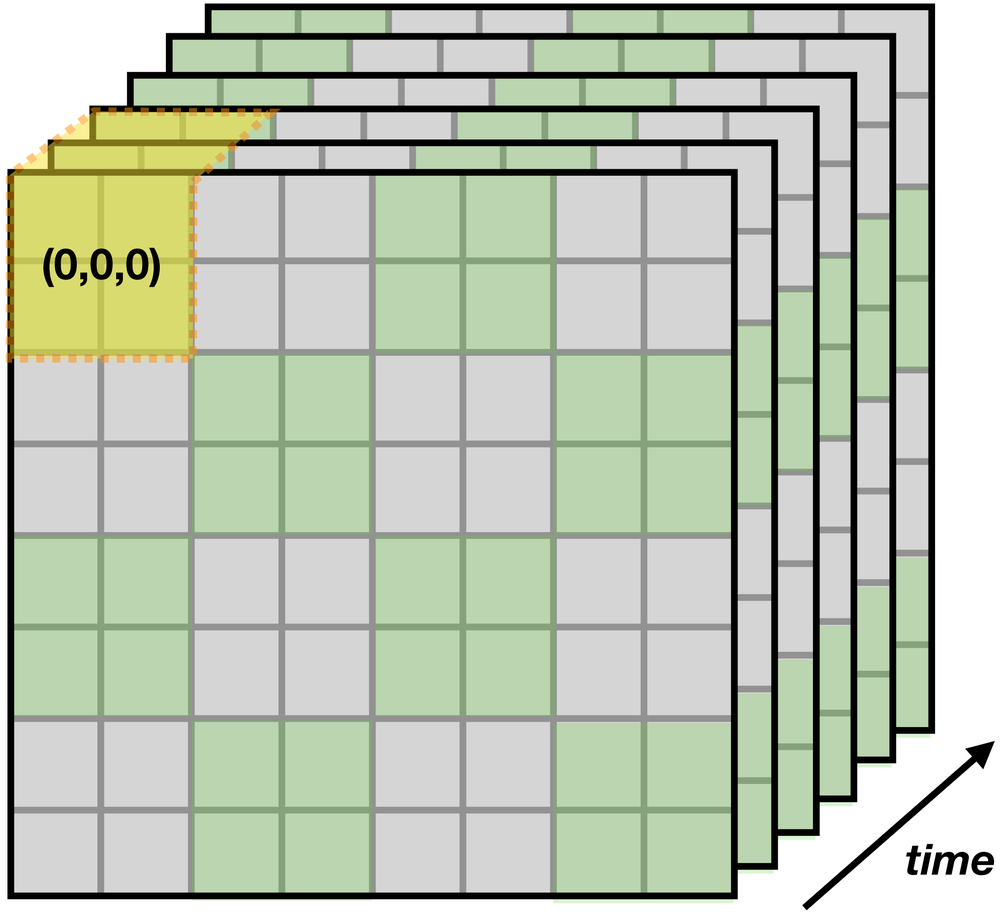
Let $\mathbf{O} = \mathbf{P}\mathbf{V} \in \mathbb{R}^{N \times d}$ ← Each output token is a weighted combination of the N projected token "values"
Where the weights are determined by the queries and keys

Tokenizing an image

Breaking an 8x8 image into 4x4 pixel chunks



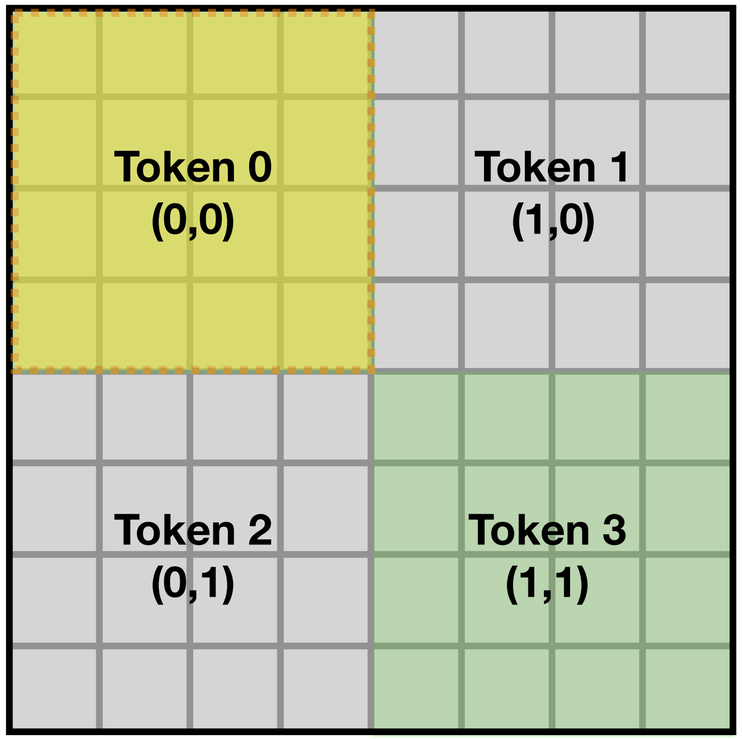
Breaking a 8x8 6-frame video into 2x2x3 pixel chunks



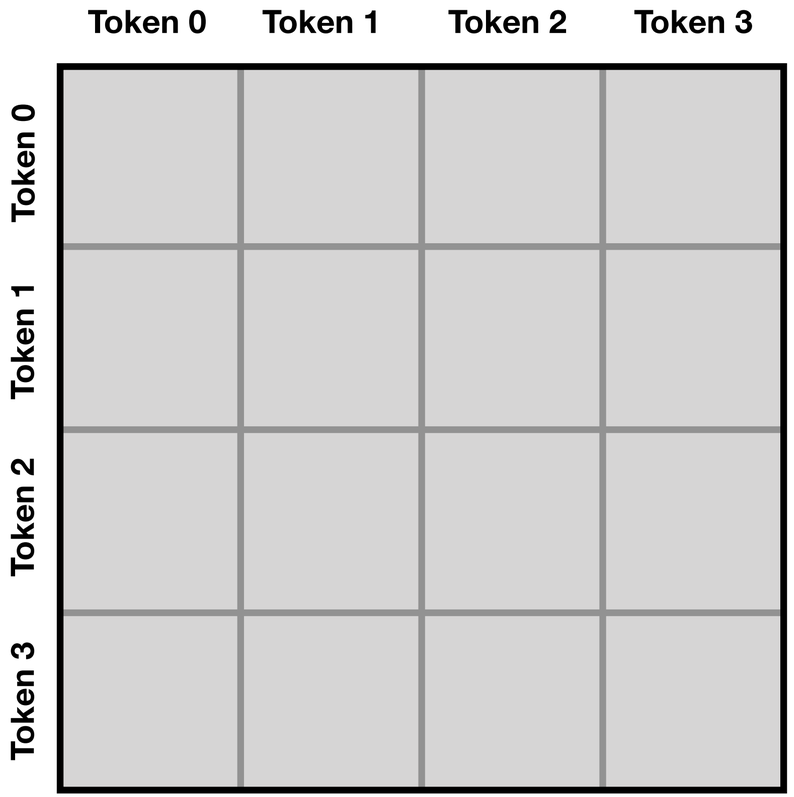
Diffusion transformer (for an 8x8 image)

- 8x8 image partitioned into four 4x4 pixel chunks (tokens)
- Yields a 4x4 attention matrix in a transformer

Breaking an 8x8 image into four 4x4 pixel chunks

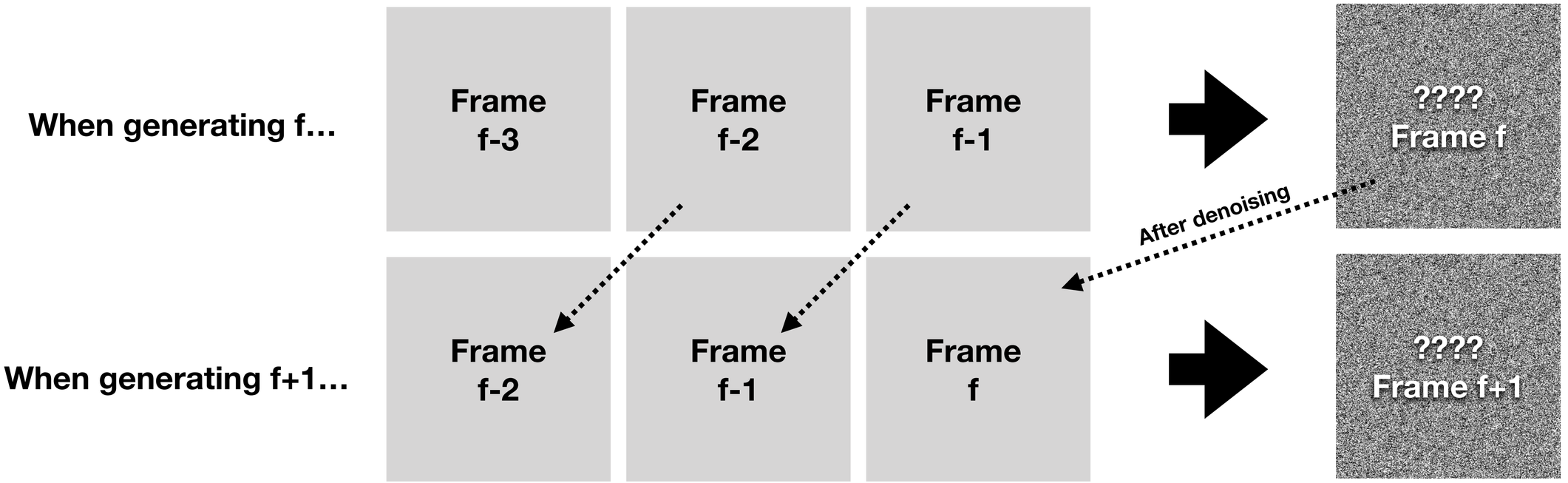


Attention matrix in the DiT
(All tokens attend to all others)



Causal, autoregressive video

- Let's imagine that we have a video with 8x8 pixel frames
- And we want to condition next frame generation based on the last 3 frames
- Wish to use diffusion to generate frame *f*
- Let's assume that tokens are 4x4x1 pixels (so there are four tokens per frame)

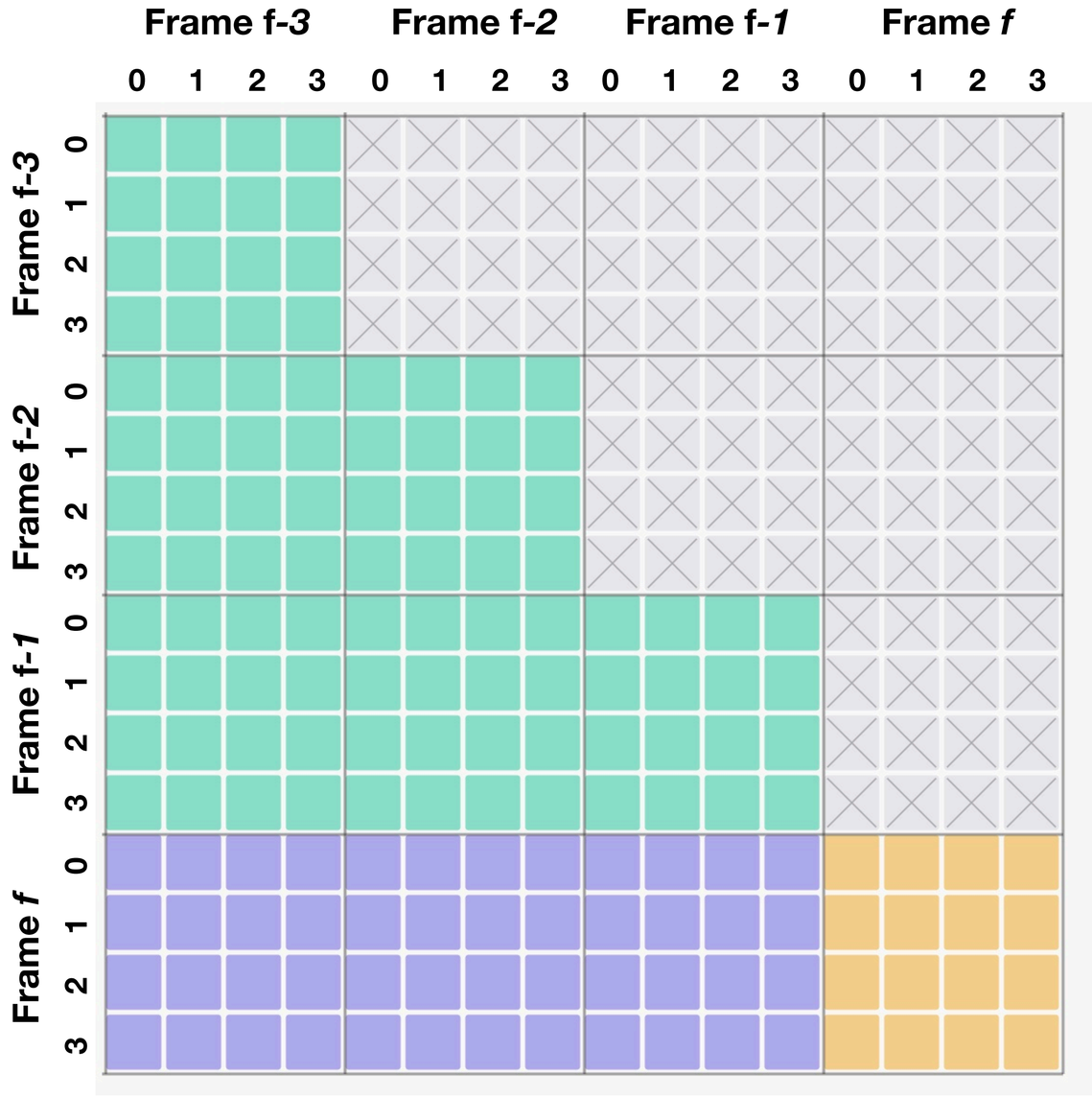


* Note: I'm using *f* to indicate frame number since *t* is used for the diffusion "step".

Casual, autoregressive video

- To generate each new frame, run T diffusion steps
- At a given diffusion step *t*, when producing frame *f*, the model is conditioned on
  - The noisy tokens from frame *f* at step *t*+1
  - The "clean" (denoised) tokens from frames f-3, f-2, f-1

- Lets look at the attention score matrix…
  - The tokens for frame *f* attend to all other tokens from frame *f*, and all tokens from prior frames



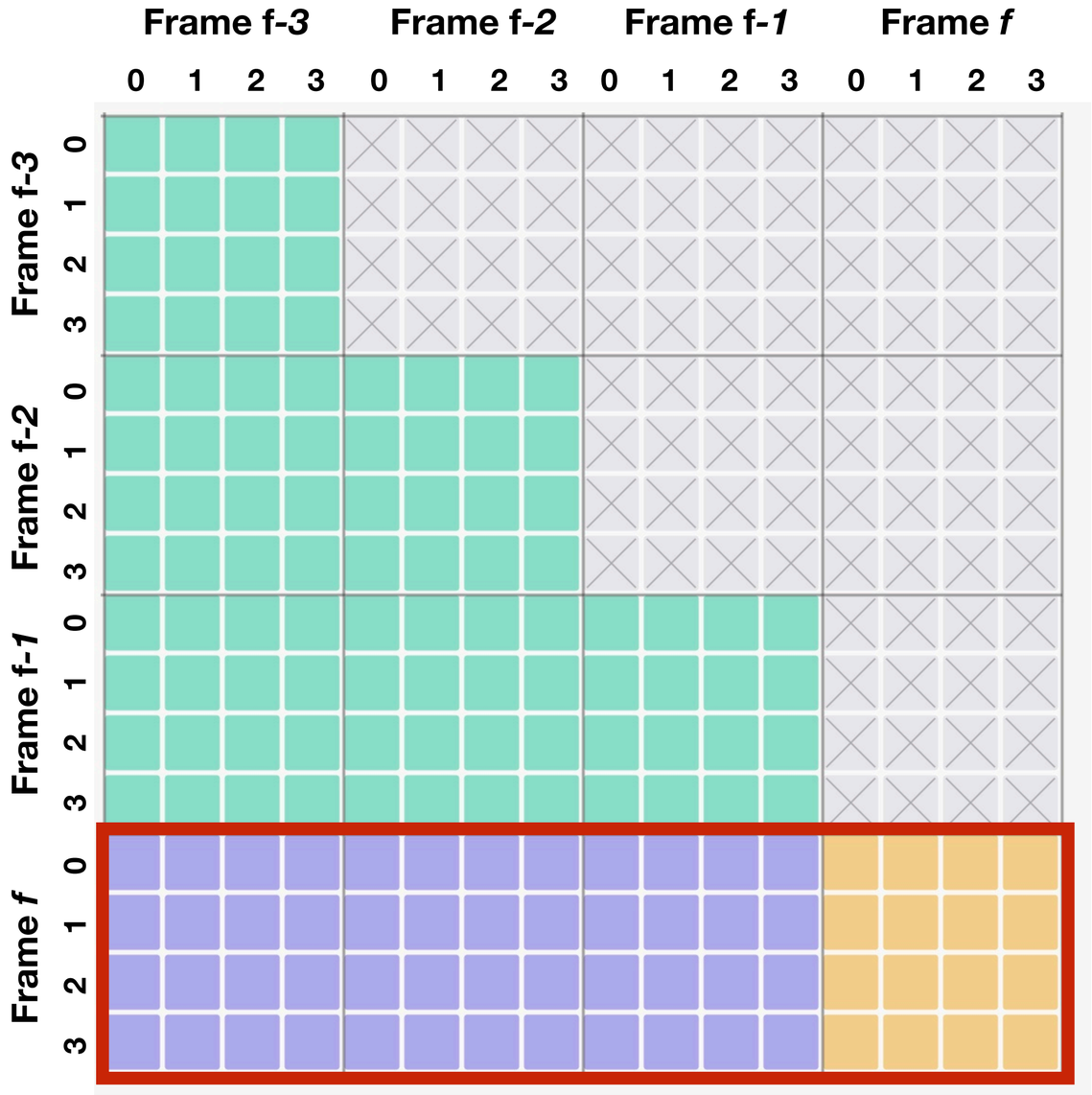
Casual, autoregressive video

- But in practice, the "keys" and "values" for the final denoised tokens for prior frames f-3, f-2, f-1 were computed in the past.
  - So only have to project the four noisy tokens

- And we only need to compute scores for the bottom four rows of the matrix!

```
Q_noisy, K_noisy, V_noisy = project(noisy_tokens)  # (4, d)
Let V = [old 12 keys | V_noisy]
Let K = [old 12 keys | K_noisy]

S = Q_noisy @ K.T                 # (4, 16)
P = softmax(S, dim=-1)            # (4, 16)
output = P @ V                    # (4, d)
```



Summary of performance optimizations on last slide

- We only had to project the noisy tokens from the current frame
  - Keys and values from tokens from prior frames are needed, but were cached from prior generations (“key-value cache”)

- We only computed rows of the score matrix for the tokens in the current frame

- We only computed the new tokens for the tokens in the current frame (multiplication of scores with value matrix)

Execute diffusion in a compact latent space

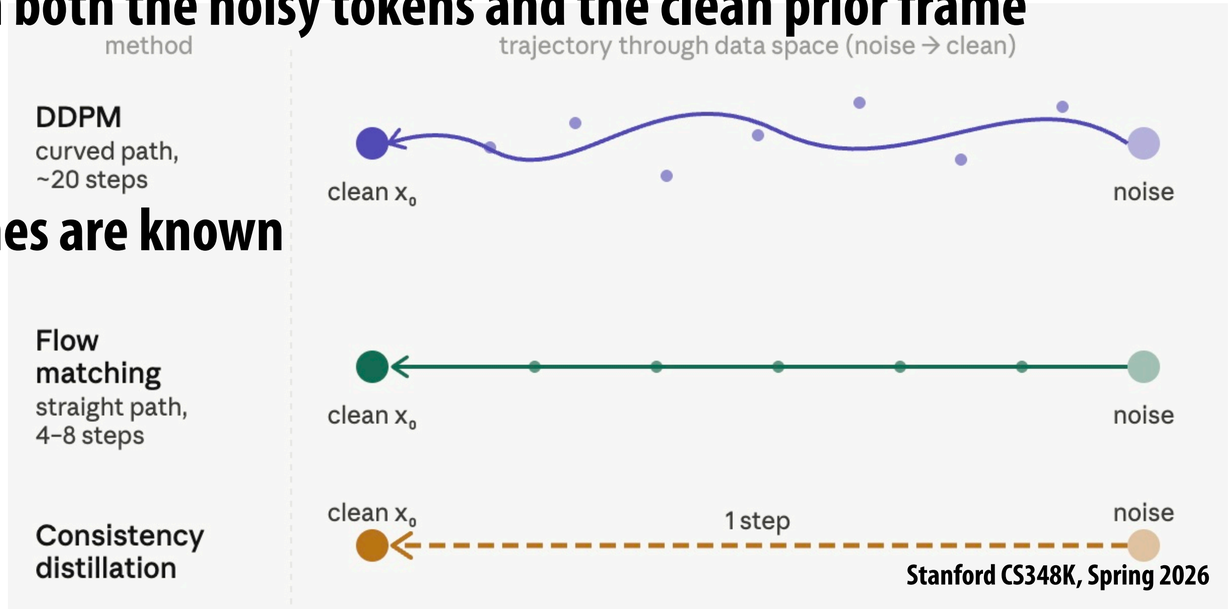
- In practice, systems to not directly turn pixels into tokens
- Learn a compact latent representation for video using a VAE (variational autoencoder), then perform diffusion on tokens represented in the latent representation
- Example:
  - Consider 1024 x 1024 x rgb video and 4x4 pixel tokens, and a 3-frame history.
  - That's 4 x 1024 x 1024 / (4 x 4) = 262K tokens… a huge score matrix!!!
- A common latent structure decimates by 8x in X and Y, and 4x in time, and uses 16 latency channels, so a video is that 1024 x 1024 x 128 frames x RGB becomes a 128x128x32x16 tensor
  - That's a 48x compression ratio
  - Now consider tokens that are 4x4x1 in latent space with a 1 token history (4-frames of history information): that's 32x32x2 =2,048 tokens going into the transformer

Reducing the number of diffusion steps

- DDPM diffusion solver must take many small steps
  - This stems from its formulation of denoising as predicting the mean and covariance of a gaussian distribution. (Gaussian approximation holds with small steps)

- Widely used alternative: flow matching objective: learn a model that given x_t, predicts a velocity vector pointed towards x_0
  - Generally allows for fewer steps to denoise since it heads "straight toward" clean tokens

- Even better: Learn to take a "huge step", conditioned on both the noisy tokens and the clean prior frame tokens: $x_{0,f} = \text{step}(x_{1,f}, x_{0,f-1}, x_{0,f-2}, x_{0,f-3})$
  - Called "consistency distillation"
  - Idea: denoising a frame is a easier when recent frames are known



Other optimizations

- Systems optimizations: efficient video generation papers also perform standard inference optimizations, such as:
  - Use of low-precision math
  - Efficient implementations of fused attention (e.g., FlashAttention)

- Also explorations into using recurrent models (see "state space models" instead of transformers to address the challenge of making history window longer)

Course Recap

A few things to walk away with

- Visual/spatial computing spans some of the most exciting applications of AI and CS today
  - Generative AI for images/videos/agents
  - Robotics / autonomous vehicles
  - New mediums of digital entertainment

- There's a lot of type out there! Understanding the pros and cons of the technologies in this space will help you place good career bets and perform impactful work

How to keep up?

- This space is moving fast! (It's hard)
- Follow some of our guest speakers or the companies that they work for, since they are active participants?
- Take a look at conference proceedings like Kesen's siggraph page or daily digests like HuggingFace's "Daily Papers" by AK
- Try to find a project to help out with in a research lab at Stanford where we talk about these topics every day (graphics, vision, robotics)

SIGGRAPH 2026 papers on the web

Page maintained by Ke-Sen Huang. If you have additions or changes, send an e-mail.

Information here is provided with the permission of the ACM

Note that when possible I link to the page containing the link to the actual PDF or PS of the preprint. I prefer this as it gives some context to the paper and avoids possible copyright problems with direct linking. Thus you may need to search on the page to find the actual document.

ACM Digital Library: ACM Transactions on Graphics (TOG) Volume 45, Issue 4 (July 2026) Proceedings of ACM SIGGRAPH 2026

SIG/TOG: Journal Paper for presentation at SIGGRAPH 2026

SIG: Conference Paper for presentation at SIGGRAPH 2026

TOG: Selected ACM TOG Paper for presentation at SIGGRAPH 2026

ACM Digital Library (DOI) Link for the paper **Paper Abstract** **Author Preprint** **Paper Video**

Paper Presentation **Paper Images** **Paper Data** **Demo Program or Source Code** **Related Links**

Changelog

(Conditionally) Accepted Papers

Gaussian Point Splatting **(SIG/TOG)**
Joris Rijsdijk, Christoph Peters, Michael Weinmann, Ricardo Marroquim (Delft University of Technology)

Matern Noise for Triangulation-Agnostic Flow Matching on Meshes **(SIG/TOG)**
Tianshu Kuai (University of Montreal & Mila), Arman Maesumi, Daniel Ritchie (Brown University), Noam Aigerman (University of Montreal & Mila)

Manifold k-NN: Accelerated k-NN Queries for Manifold Point Clouds (SIG/TOG)
Pengfei Wang, Qinghao Guo, Haisen Zhao, Shiqing Xin (Shandong University), Shuangmin Chen (Qingdao University of Science and Technology), Changhe Tu (Shandong University), Wenping Wang (Texas A&M University)

DeepMill++: Neural Guidance Meets Rasterization for Efficient Accessibility Analysis **(SIG)**

A few things to walk away with

- Ability to define problems explicitly (defining goals and constraints) is a very universal clear thinking skill
  - It helps illuminate how a system should be evaluated
  - If helps suggest a set of possible solutions

- Demand it from yourself, and your collaborators/coworkers/colleagues/teammates

- Give the speed at which our field is moving, we need folks to be demanding clear thinking more than ever!

Thanks!

- Thanks for being a great class!
- Looking forward to seeing your projects on Tuesday!