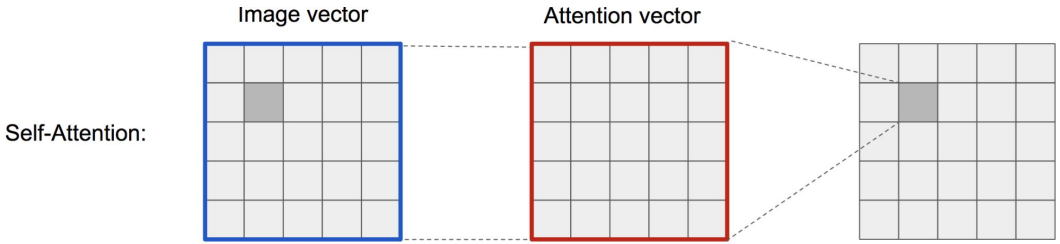
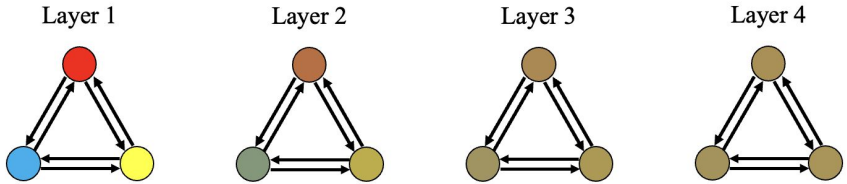
Generative Modeling as Operator Design: From Diffusion to Attention

Atieh Armin

Overview

- Evolution of generative models
- Limitations of different models
- Graphs make it harder
- Goal: an attention-based operator with built-in structure and stability





01 Markov Chains & Langevin Dynamics

02 Score-based & Diffusion Models

03 Transformers & Attention Mechanism

04 GNNs & Graph Attention Networks

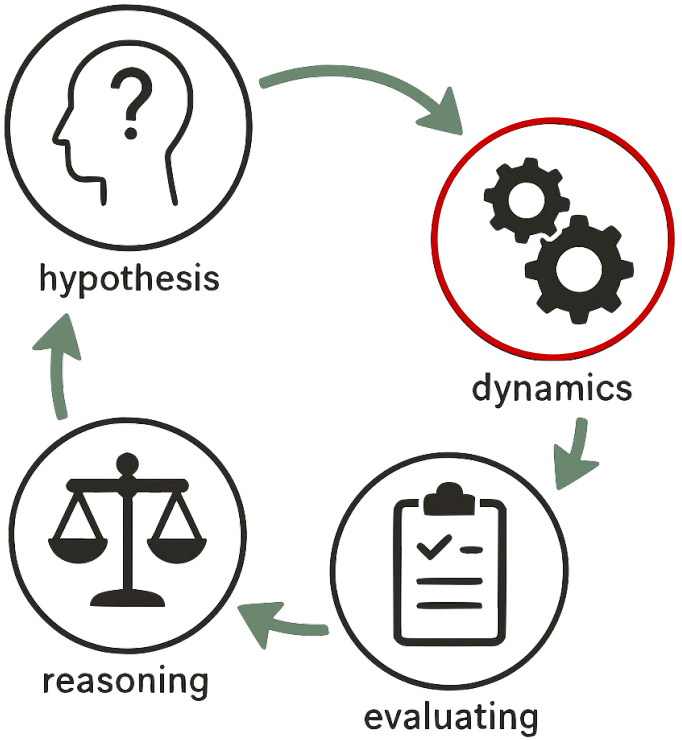
05 Bridging Dynamics and Attention

06 Applications & Future Work

01 – Markov Chains & Langevin Dynamics

Generative Modeling

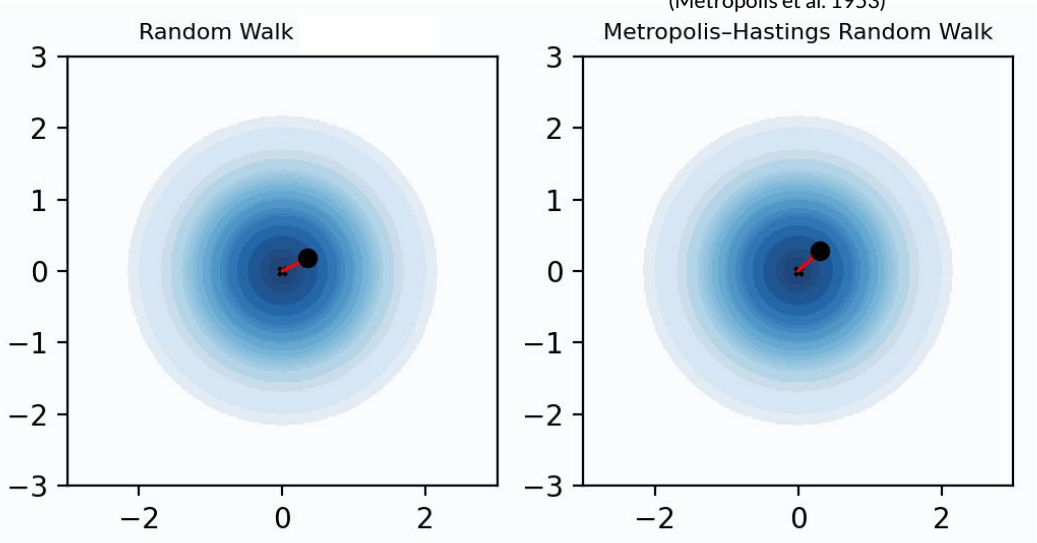
A generative model tries to imitate how real data is made so it can generate similar examples.



Markov Chains Monte Carlo (MCMC) (Hastings 1970)

- State evolves using a transition kernel $P(x' \mid x)$
- Desired distribution is stationary under P
- Detailed balance:
  - $\pi(x)\,P(x' \mid x) = \pi(x')\,P(x \mid x')$
  - Ensures reversibility + convergence (Meyn & Tweedie 2012)
- Spectral gap controls mixing efficiency (Levin & Peres 2017)

(Metropolis et al. 1953)

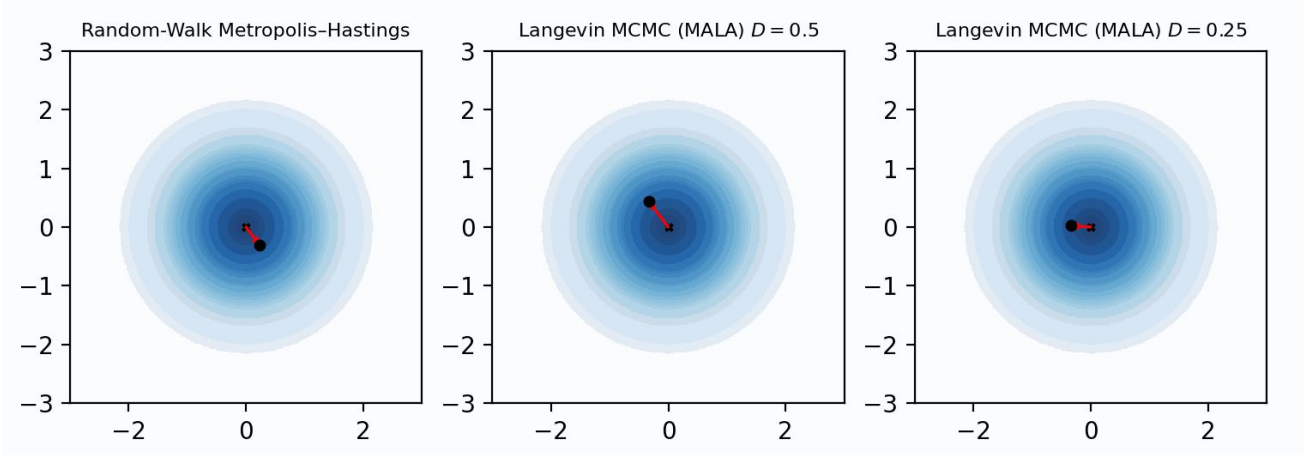


Langevin Dynamics

- Adds gradient information
- Deterministic drift + Stochastic noise

$$dX_t = -\frac{1}{\gamma}\nabla U(X_t)\,dt + \sqrt{2D}\,dW_t$$

- Moves toward high-density regions while exploring
- Reversible (Risken 1989) and has strong convergence guarantees (Bakry et al. 2013)



Langevin Dynamics

- Unadjusted Langevin Algorithm(ULA):

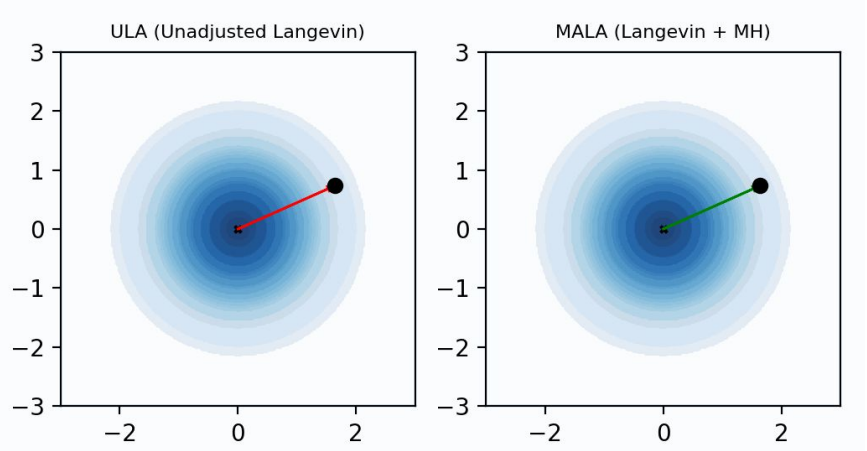
$$x_{k+1} = x_k - \frac{\eta}{\gamma} \nabla U(x_k) + \sqrt{2D\eta} \cdot \xi_k$$

- Metropolis-Adjusted Langevin Algorithm (MALA) (Roberts & Tweedie 1996) adds a Metropolis–Hastings acceptance step:

$$\alpha(x, x') = min\{1, \frac{\pi(x')\ q(x \mid x')}{\pi(x)\ q(x' \mid x)}\}$$

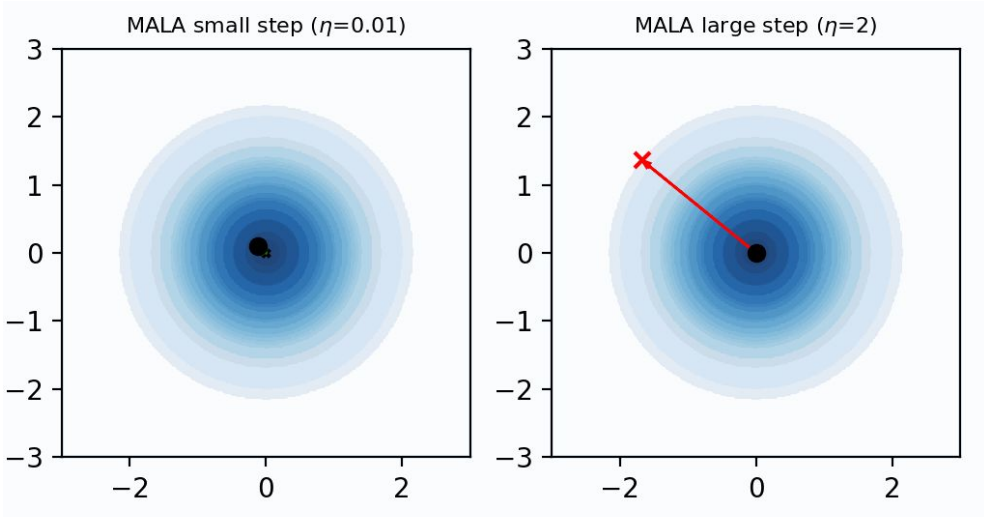
$$u \sim Unif\ (0, 1)$$

accept x' if $u \leq \alpha(x, x')$, otherwise keep x



Langevin Dynamics

- Limitations:
  - Trade-offs: step-size vs bias vs efficiency
  - Requires knowing the target distribution and its gradient



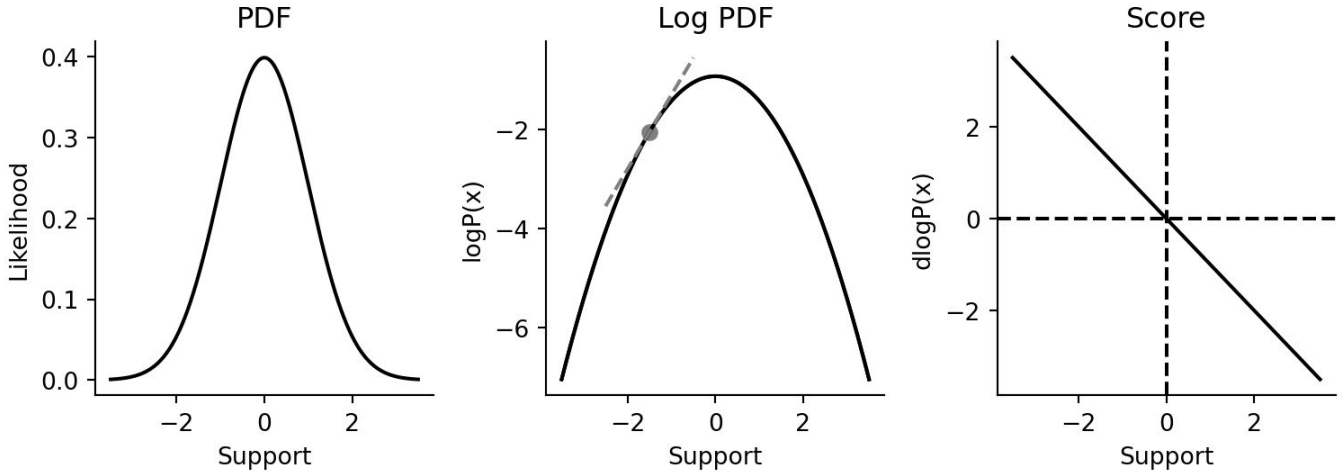
$$x_{k+1} = x_k - \frac{\eta}{\gamma} \nabla U(x_k) + \sqrt{2D\eta} \cdot \xi_k$$

02 – Score based & Diffusion Models

Score-Based Modeling

- Learn the score function directly from data

$$s(x) = \nabla_x \log p(x)$$



Score-Based Modeling

Score matching (Hyvärinen 2005)

- Model's score: $s_\theta(x) = \nabla_x \log p_\theta(x)$
- Data's score: $s_{\text{data}}(x) = \nabla_x \log p_{\text{data}}(x)$
- Objective:

$$J(\theta) = \mathbb{E}_{p_{\text{data}}}\left[\frac{1}{2} \| s_\theta(x) - s_{\text{data}}(x) \|^2\right].$$

- Limitations:
  - It requires either the true data score or the second order derivatives

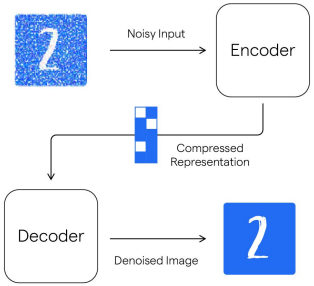
Denoising Score Matching (Vincent 2011)

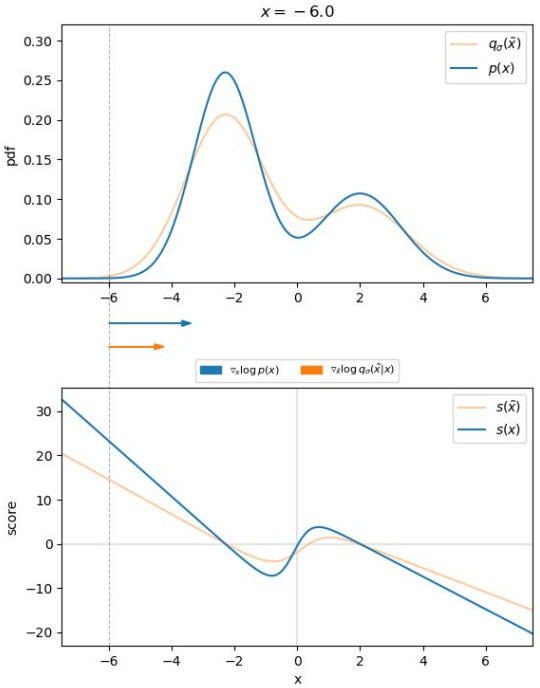
- Denoising Autoencoders:
  - Learn a model that denoises a corrupted sample to get back to the original sample
- Score-matching view:
  - Defining the score-matching objective for the corrupted dataset:

$$J(\theta) = \mathbb{E}_{q_{\sigma(\tilde{x})}}\left[\frac{1}{2} \| s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}) \|^2\right]$$

  - Vincent shows that this objective can be considered as equivalent to:

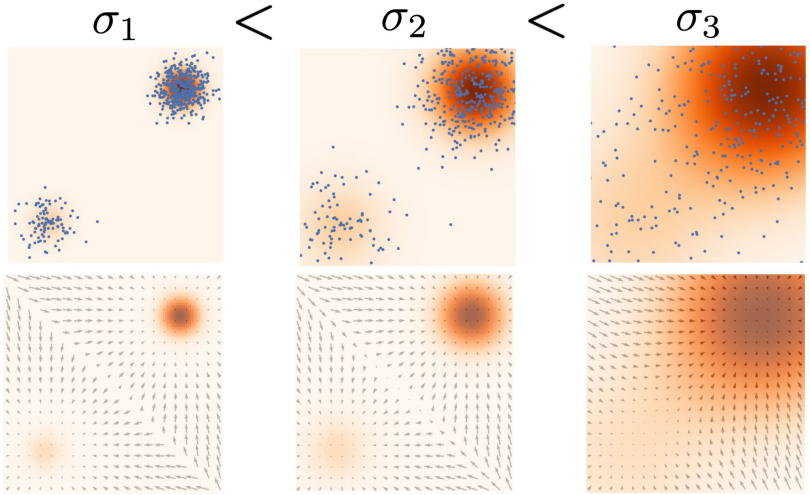
$$J(\theta) = \mathbb{E}_{q_{\sigma(\tilde{x},x)}}\left[\frac{1}{2} \| s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) \|^2\right], \textit{where } \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) = \frac{1}{\sigma^2}(x - \tilde{x}) .$$





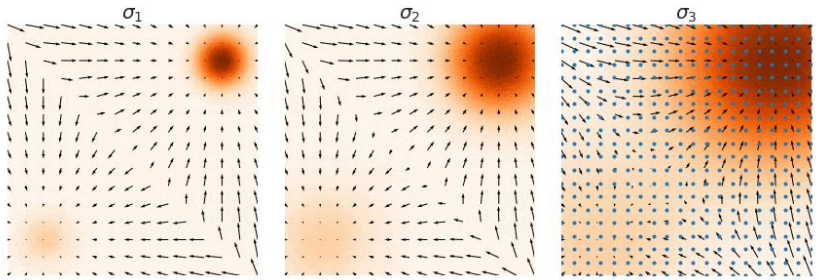
Noise-Conditional Score Networks (Song & Ermon 2019)

1. **Learning NCSN via Score Matching:** Train the score network on perturbed data (on multiple noise levels) using denoising score matching => $S_\theta(x, \sigma)$



2. **NCSN Inference via Annealed Langevin Dynamics:**

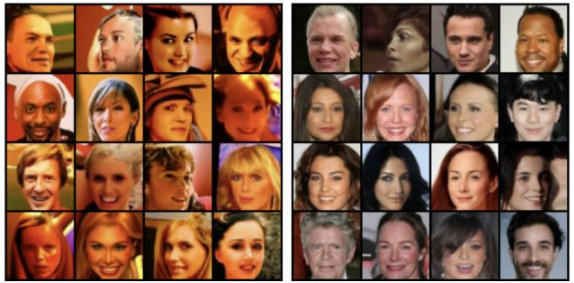
$$x_t = x_{t-1} - \frac{\eta}{2} S_\theta(x, \sigma) + \sqrt{\eta} \cdot \xi_k$$



Noise-Conditional Score Networks (Song & Ermon 2019)

Limitations:

- Sampling is very slow (many Langevin steps)
- Limited to low-resolution images & unstable training (Song & Ermon 2020)

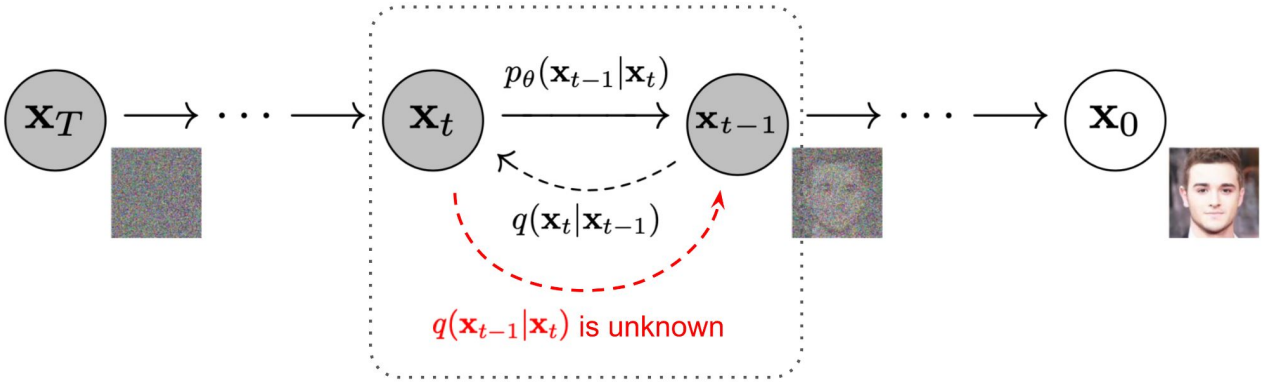


(a) NCSN (b) NCSNv2

Figure 10: Uncurated samples from NCSN (a) and NCSNv2 (b) on CelebA 64×64.

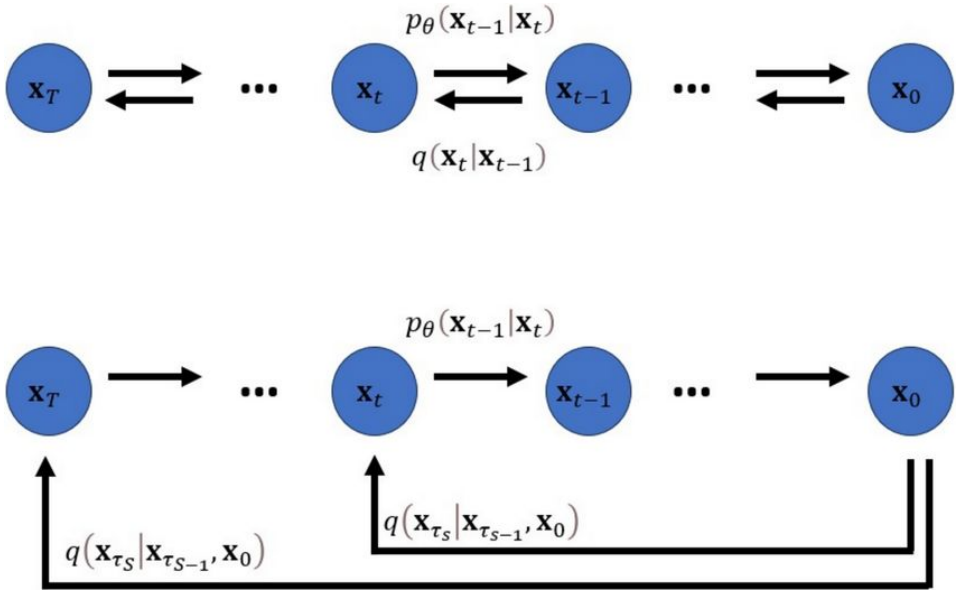
Denoising Diffusion Probabilistic Models (Ho et al. 2020)

- Add noise gradually and learn to predict the added noise
- After training:
  - Start with random noise
  - Run the reverse Markov chain (denoising steps)
- Still iterative and expensive

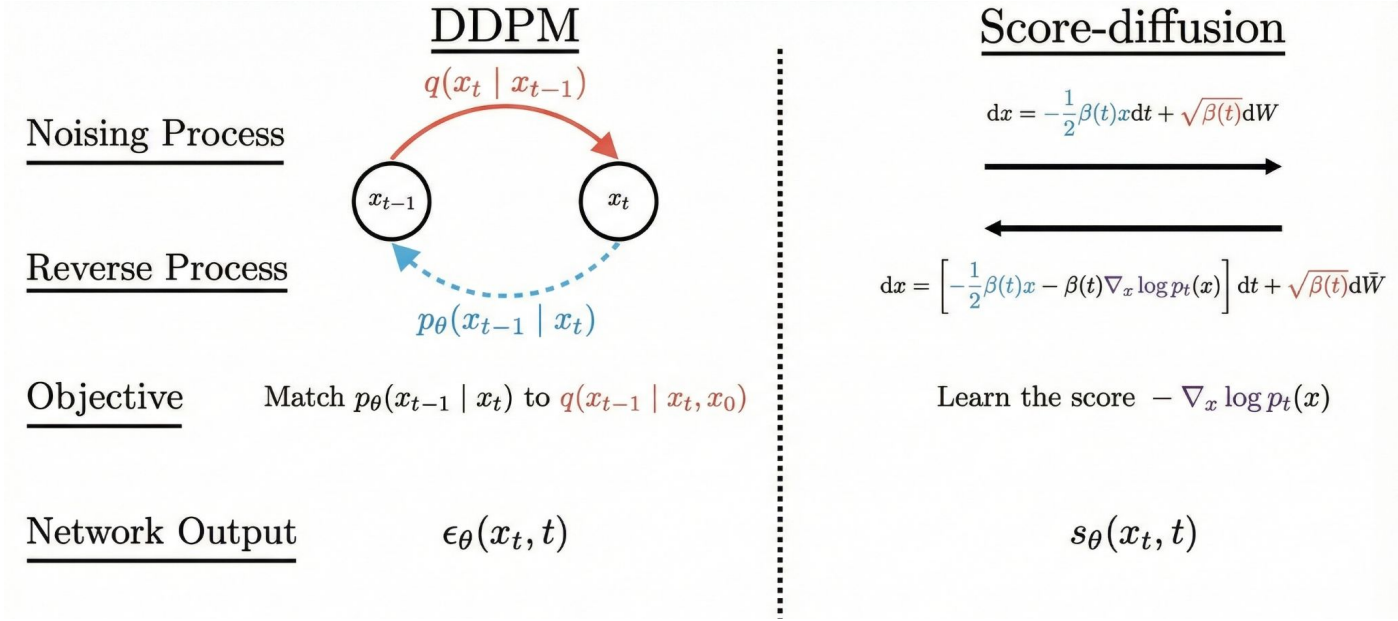


Denoising Diffusion Implicit Models (Song et al. 2020)

- In DDPM, the forward process is a Markov chain!
- In DDIM, they consider the forward process as a non-Markovian process
  - Allowing you to skip steps in the denoising process, not requiring all past states to be visited before the current state.
- Changes sampling step
  - From this noisy image, jump directly to the version that would correspond to an earlier (cleaner) noise level
- Preserves quality while allowing speed–quality tradeoff



Score-Based Diffusion Models (Song et al. 2020)



Score-Based Diffusion Models

- This let to the first generation of high-resolution images from score-based models.
- However, they remained computationally expensive + requiring a lot of steps

Score-Based Models as Operators

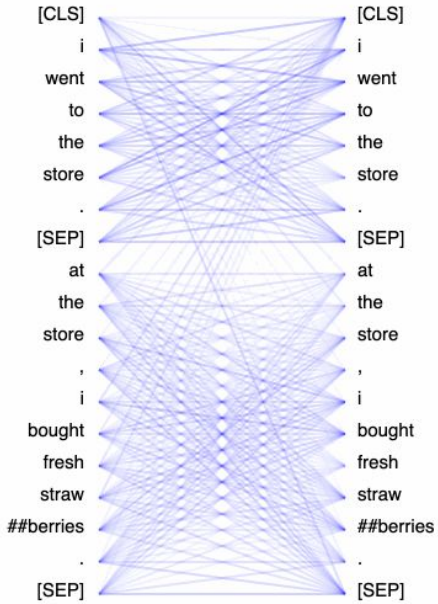
- Learn a family of operators that gradually transform noise → data
- Training: estimate the score
- Inference: reverse-time process with many small operator steps
- Generation: long sequence of local updates
- Limitations:
  - local steps struggle in low-density or ambiguous regions
  - sampling requires hundreds of iterations (slow)
- Motivates the need for more global, expressive, and efficient operators

03 – Transformers & Attention Mechanism

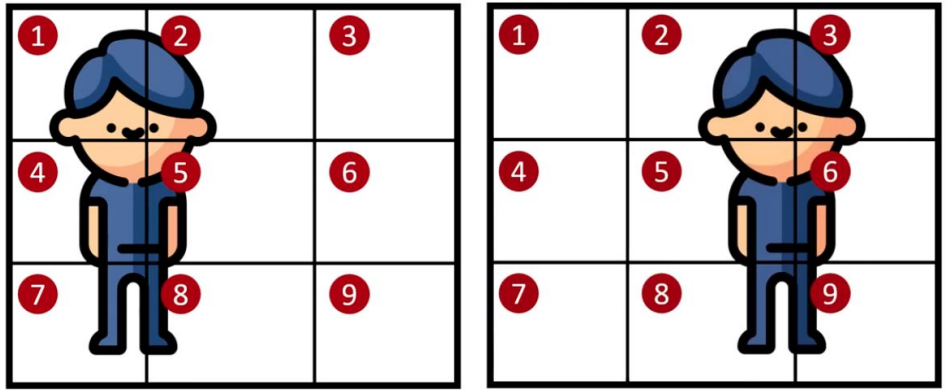
Attention is All You Need!

(Vaswani et al 2017)

- Attention is a global, content-adaptive operator
- Allows long-range interactions in one step
- Strength: expressiveness + parallelism

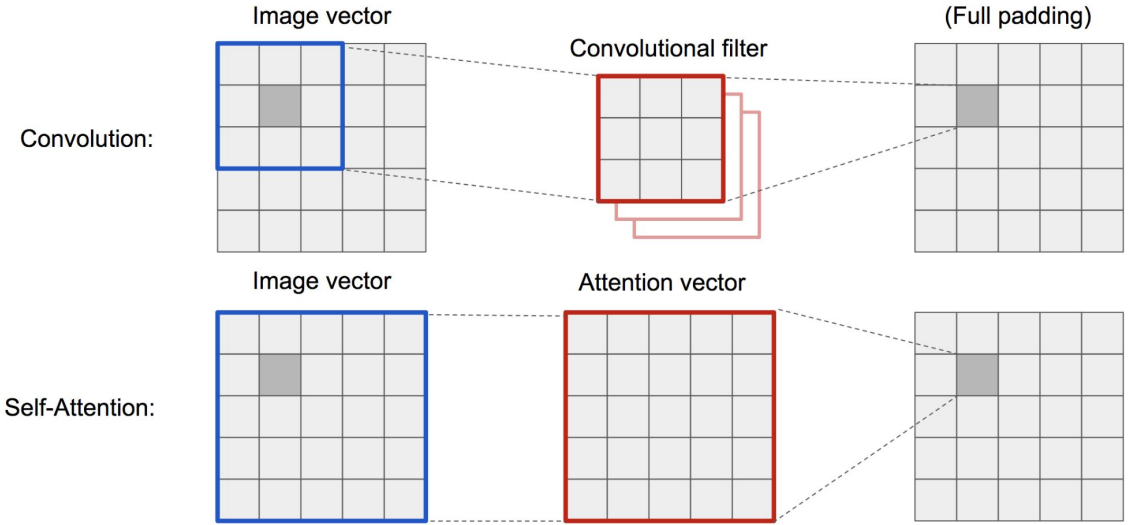


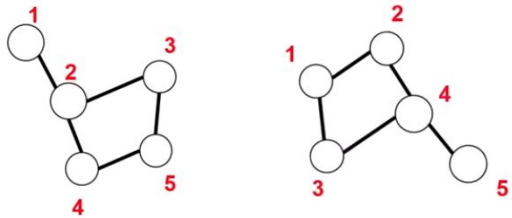
Is the Attention All We Need?



Limitations:

1. Lack of spatial bias / locality

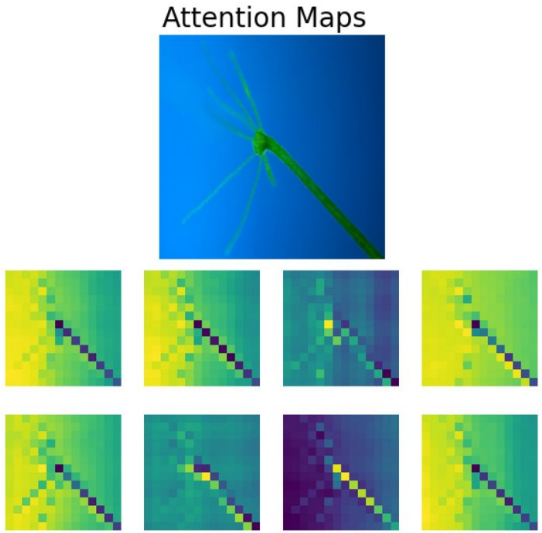


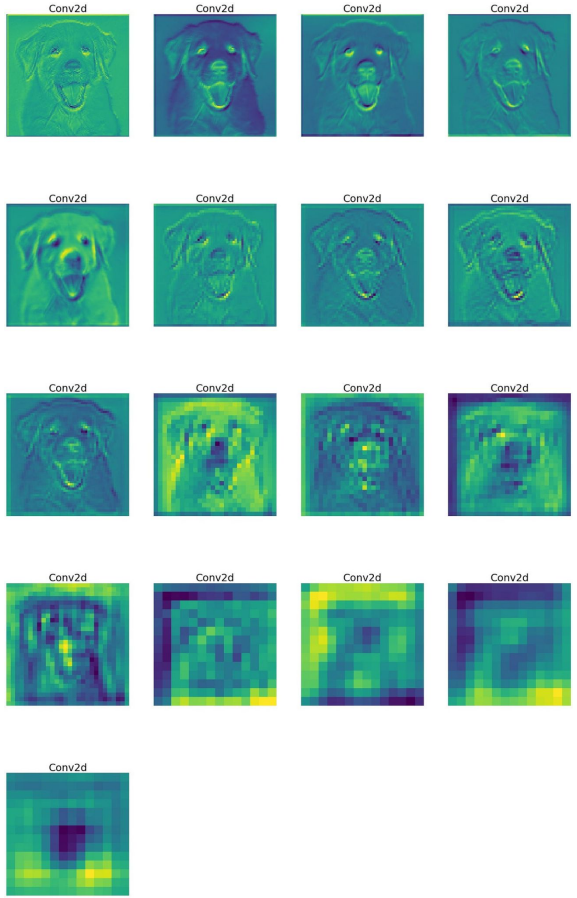


Is the Attention All We Need?

Limitations:

2. No inherent multiscale or frequency awareness



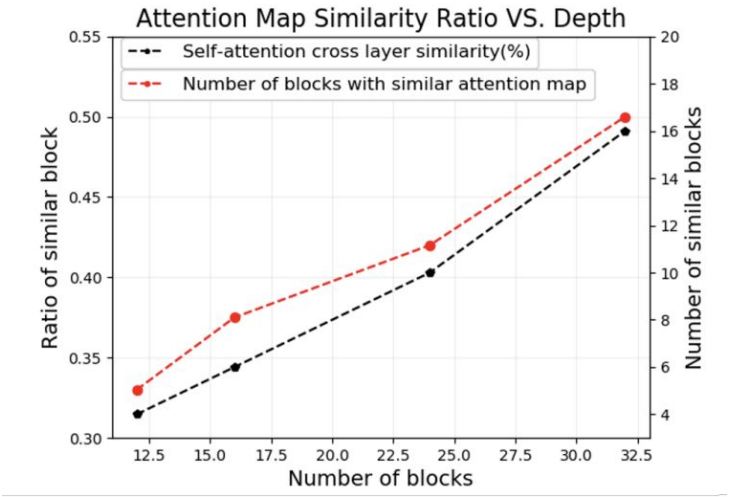


CNN Feature Maps

Is the Attention All We Need?

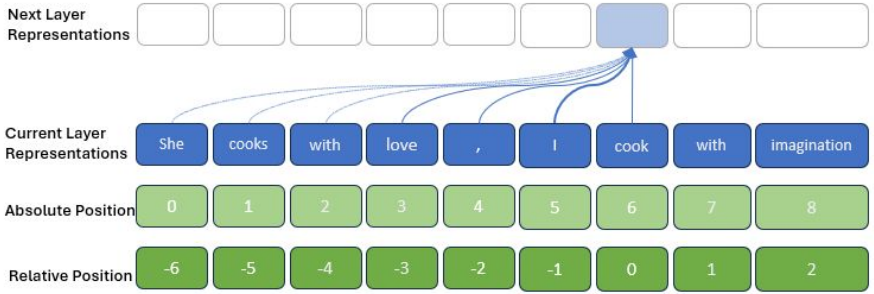
Limitations:

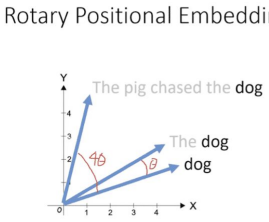
3. Unconstrained global interactions → premature mixing / oversmoothing-like behavior

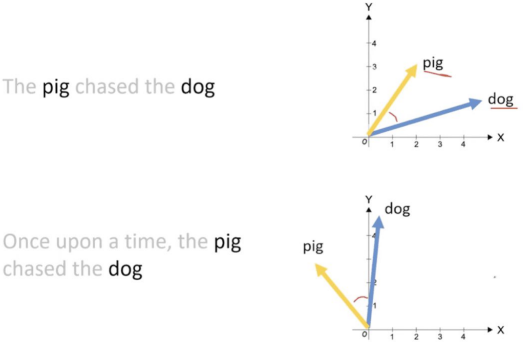


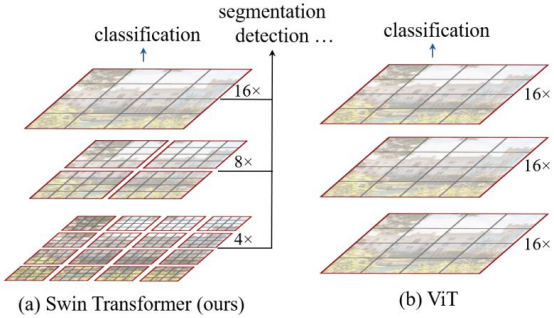
Solutions

- Relative Positional Embeddings (Shaw et al. 2018)
- Rotary Positional Embeddings (Su et al. 2024)
- Swin Transformers (Liu et al. 2021)









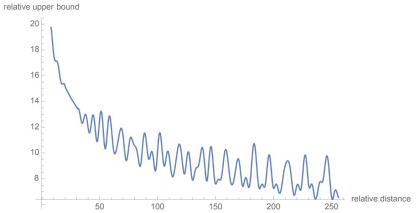
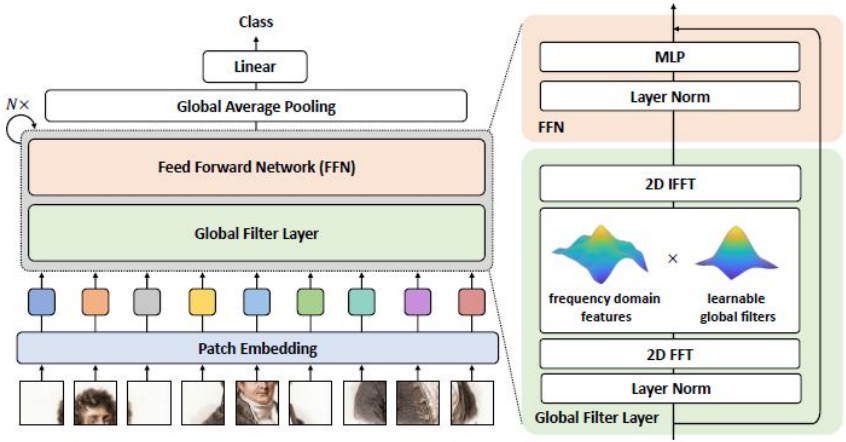


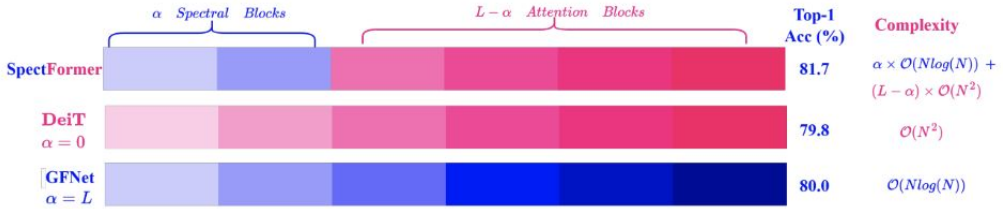
Figure 2: Long-term decay of RoPE.

Solutions

- GFNet (Rao et al. 2021)



- Spectformer (Patro et al. 2025)



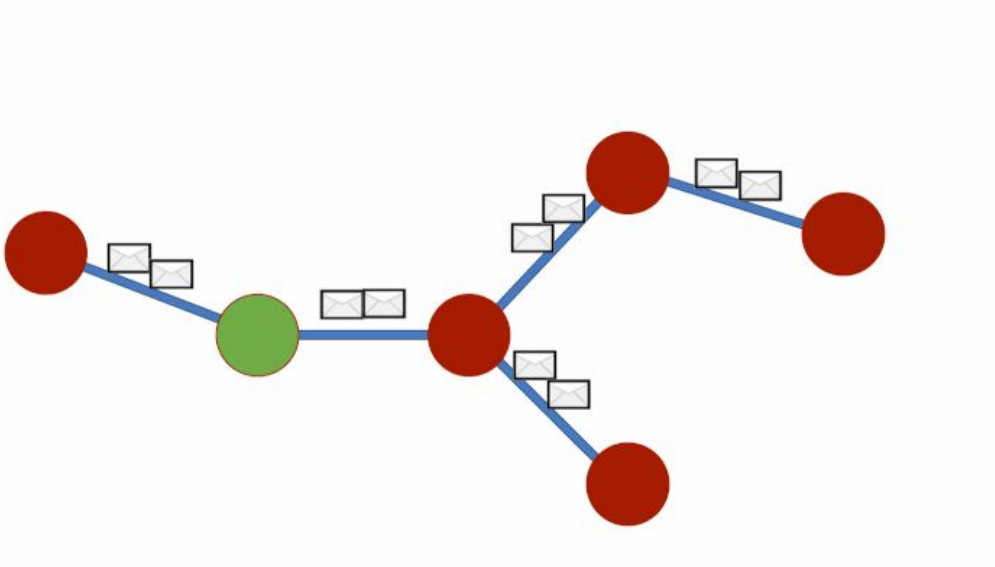
Solutions

- ➔ All of these, augment attention or replace it but don’t regulate its internal operator!
- ➔ Can we reshape the attention kernel itself to behave like a well-structured operator?

04 – GNNs & Graph Attention Networks

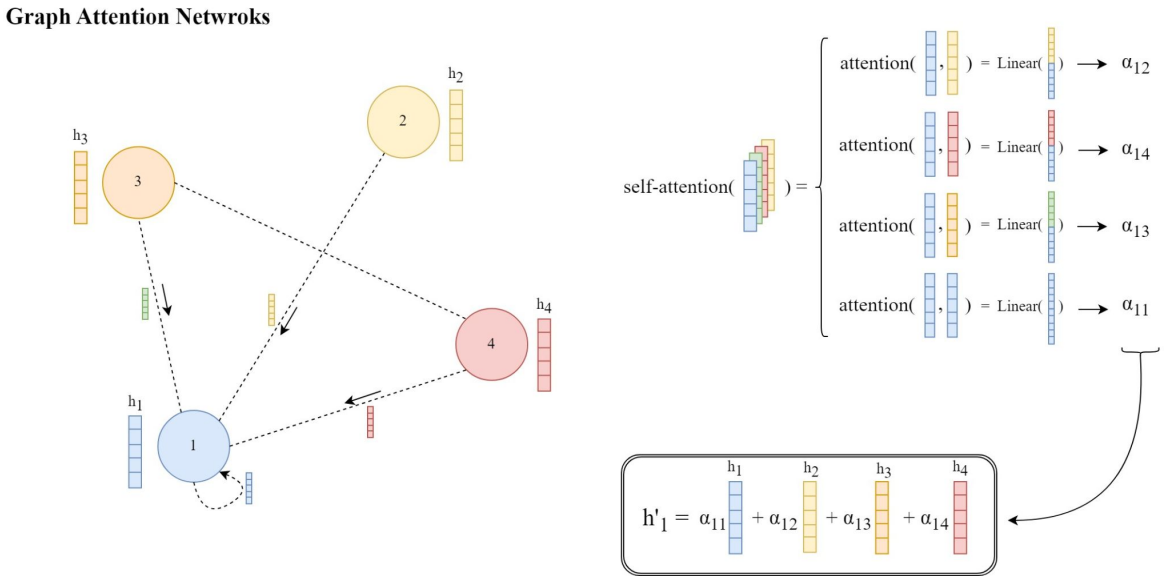
Graph Convolutional Networks (Kipf et al. 2017)

- Message passing gradually mixes node features across the graph
- Advantages: locality, geometry-awareness
- Limitations: over-smoothing (Oono & Suzuki 2020)



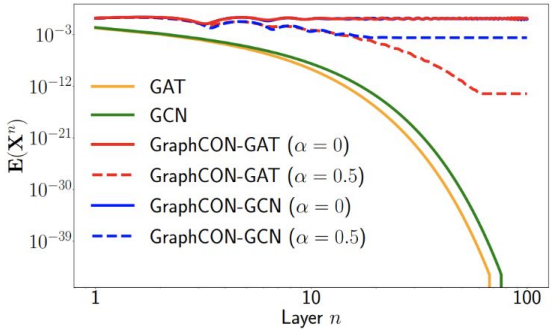
Graph Attention Networks (Velickovic et al. 2018)

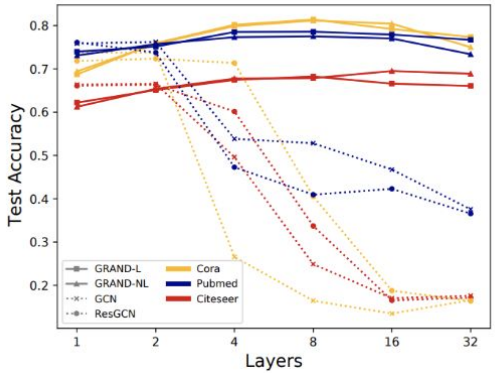
- Introduce learned, data-dependent attention over neighbors
- Strong performance on node/graph classification
- Can still suffer from over-smoothing as depth increases (Wu et al. 2023)



What's the Problem?

- Missing ingredient: physics-inspired operators that combine local propagation, stability, and long-range interaction
- Recent works have begun to explore this direction:
  - GRAND (Chamberlain et al. 2021)
  - GraphCON (Rusch et al. 2022)





05 – Bridging Dynamics and Attention

Central Hypothesis

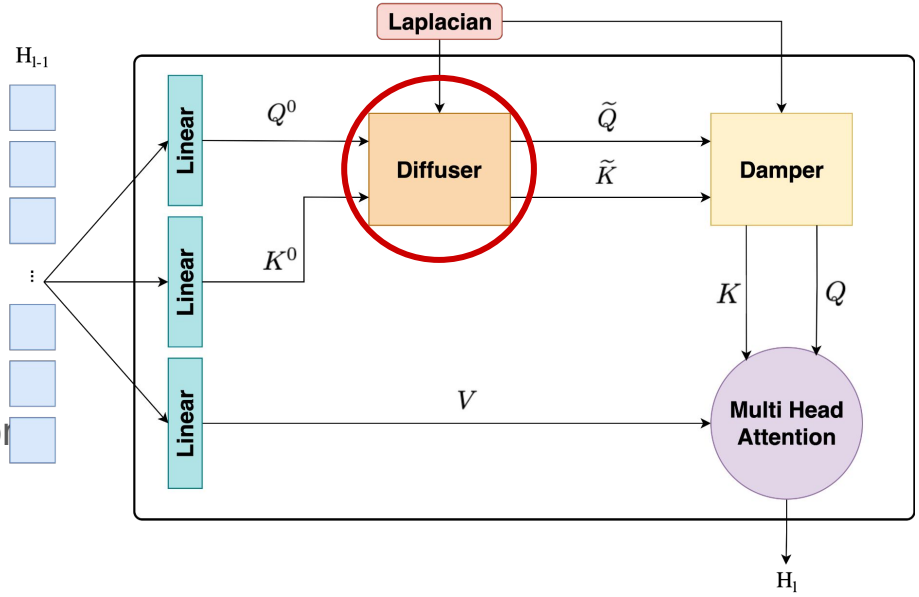
Attention can internalize principled dynamics:

- Embed physically inspired operators into the attention kernel
- Make attention:
  - geometry-aware
  - spectrally regularized
  - energy-controlled
  - stable across depth

Proposed Operator

The attention kernel will be a diffusion-regularized spectral operator, incorporating:

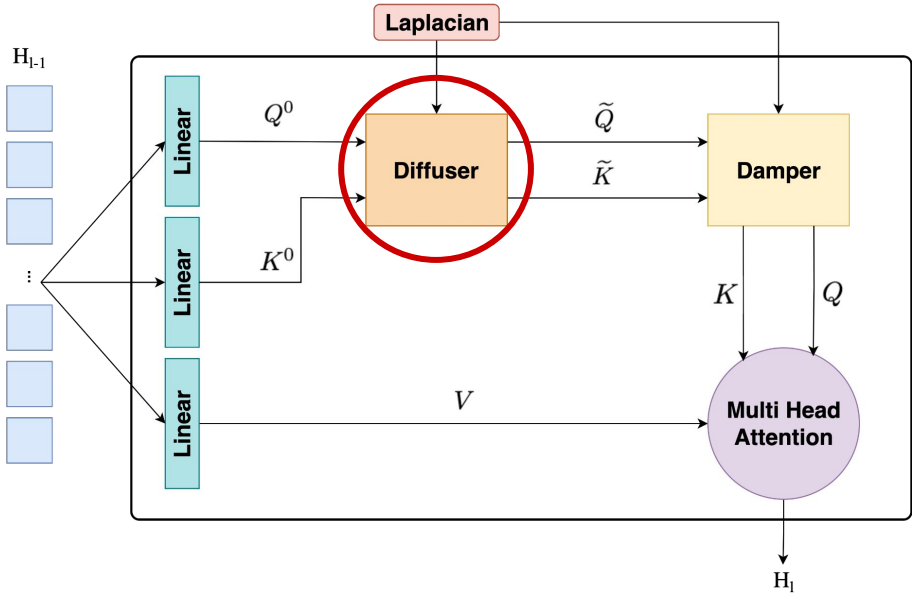
1. Diffusion-Based Propagation
    - Use graph Laplacian to anchor interactions
    - Encourages local coherence
    - Integrates message passing into attentior



Proposed Operator

The attention kernel will be a diffusion-regularized spectral operator, incorporating:

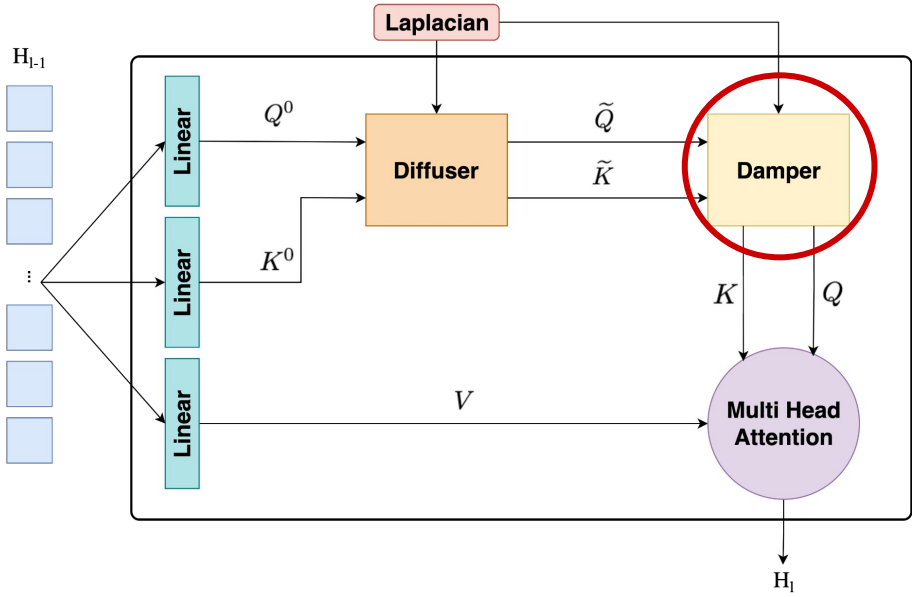
2. Harmonic/Spectral Bias
    - Filter high frequencies
    - Promote smooth, coherent signals
    - Avoid noise amplification
    - Spectrally shapes the attention matrix



Proposed Operator

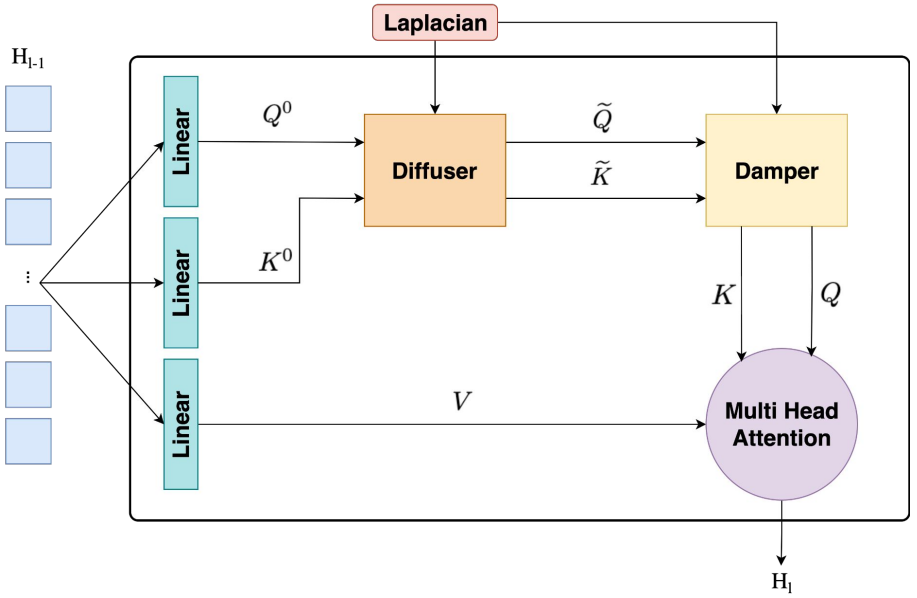
The attention kernel will be a diffusion-regularized spectral operator, incorporating:

3. Damping Mechanism
    - Controls energy across layers
    - Prevents attention blow-up and chaotic long-range propagation
    - Formal operator remains bounded



Operator Properties

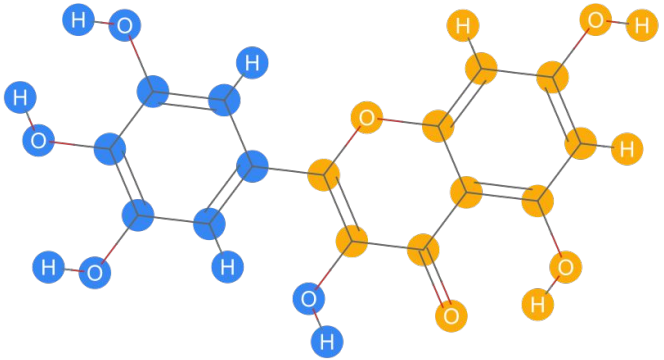
- Geometry-aware
- Spectrally controlled
- Stable across depth
- Globally expressive
- Nonlocal like attention, structured like diffusion

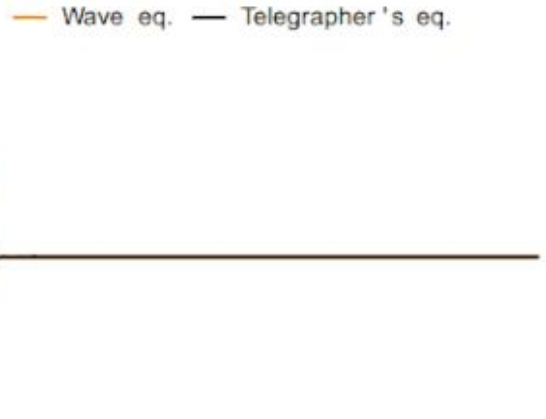


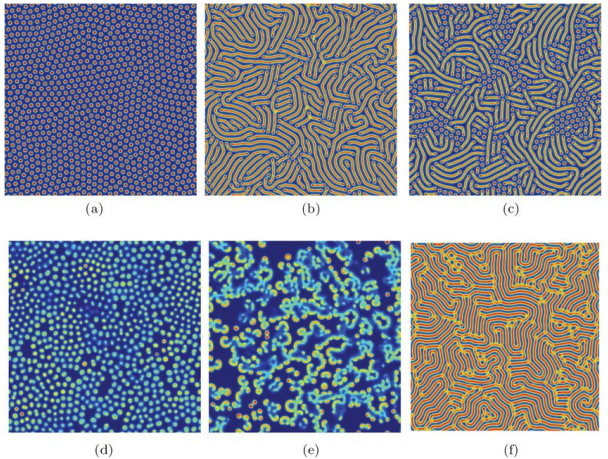
06 – Applications & Future Work

Operator Design for Different Systems

- Molecular graphs
- Reaction–diffusion systems
- Wave/telegraph dynamics







Fast Approximations

- Physically inspired operators can be costly
- Chebyshev polynomial approximations:
  - Scalable
  - Structure-preserving
  - Ideal for spectral filters

Conclusion

- Generative modeling = operator design
- Diffusion → local, stable, slow
- Transformers → global, expressive, unstable
- Graph-based models → local smoothing, over-smoothing
- Proposed direction:
  - embed dynamics inside attention
  - design inherently structured operators
  - unify global expressiveness with physical grounding

Thank you for your attention!