Lecture 4

Conditional Image Generation

MIT IAP 2025 | Jan 27, 2025

Peter Holderrieth and Ezra Erives

Sponsor: Tommi Jaakkola

Recall: So far we have focused on **unconditional** generation.

Problem: Sample from p_{data}

Train: Use e.g., the conditional flow matching objective:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{\square} \| u_t^\theta(x) - u_t^{\text{target}}(x|z) \|^2$$

$$\square = z \sim p_{\text{data}},\ t \sim \text{Unif}[0,1),\ x \sim p_t(x|z)$$

Sample: Simulate the corresponding ODE (or SDE):

$$\mathrm{d}X_t = u_t^\theta(X_t)\mathrm{d}t, \qquad X_0 \sim p_{\text{init}}$$

But what about **conditional generation?**

Today's Agenda:

1. Extend our generative modeling framework from **unconditional generation** to **conditional generation**
2. Develop **classifier-free guidance** for conditional sampling
3. Discuss **architectural choices** for the prototypical case of **image generation** and **survey current models.**
4. Guest talk by Carles Domingo-Enrich!

Part 1:
Conditional Generation and Guidance

A swamp ogre with a pearl earring by Johannes Vermeer

A car made out of vegetables.

heat death of the universe, line art

Image source: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis [1]

Unconditional: “Generate an image.”

Conditional: “Generate an image of a cat baking a cake.”

A swamp ogre with a pearl earring by Johannes Vermeer

A car made out of vegetables.

heat death of the universe, line art

Image source: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis [1]

~~**Unconditional**~~ **Unguided**: “Generate an image.”

~~**Conditional**~~ **Guided:** “Generate an image of a cat baking a cake.”

Guided Generation: What Changes?

Unguided		Guided	
Marginal probability path	$p_t(x)$	**Guided** marginal probability path	$p_t(x\|y)$
Marginal vector field	$u_t^{\text{target}}(x)$	**Guided** marginal vector field	$u_t^{\text{target}}(x\|y)$
Marginal score	$\nabla \log p_t(x)$	**Guided** marginal score	$\nabla \log p_t(x\|y)$
Model	$u_t^\theta(x)$	**Guided** model	$u_t^\theta(x\|y)$
CFM Objective	$\mathcal{L}_{\text{CFM}}(\theta)$	**Guided** CFM Objective	???

A Guided CFM Objective

Observation: For **fixed y**, we obtain the **unguided problem**, and may adapt an **unguided objective** to obtain:

$$\mathcal{L}_{\text{CFM}}^{\text{guided}}(\theta; y) = \mathbb{E}_{\square}\|u_t^\theta(x|y) - u_t^{\text{target}}(x|z)\|^2$$

$$\square = z \sim p_{\text{data}}(z|y),\ t \sim \text{Unif}[0,1),\ x \sim p_t(x|z)$$

Observation: By **varying y**, the above yields a **guided objective** for **general y**:

$$\mathcal{L}_{\text{CFM}}^{\text{guided}}(\theta) = \mathbb{E}_{\square}\|u_t^\theta(x|y) - u_t^{\text{target}}(x|z)\|^2$$

$$\square = (z, y) \sim p_{\text{data}}(z, y),\ t \sim \text{Unif}[0,1),\ x \sim p_t(x|z)$$

We may then **train** using this objective.

Guided Sampling

Algorithm 7 Guided Sampling Procedure

Require: A trained guided vector field $u_t^\theta(x|y)$.

1: Select a prompt $y \in \mathcal{Y}$, such as "a cat baking a cake".
2: Initialize $X_0 \sim p_{\text{init}}$.
3: Simulate $\mathrm{d}X_t = u_t^\theta(X_t|y)\mathrm{d}t$ from $t = 0$ to $t = 1$.

Can we do better? At least empirically, the answer is yes…

Classifier-Free Guidance

For Gaussian probability paths, it can be shown that

$$u_t^{\text{target}}(x|y) = u_t^{\text{target}}(x) + b_t \nabla \log p_t(y|x), \qquad b_t = \frac{\dot{\alpha}_t \beta_t^2 - \dot{\beta}_t \beta_t \alpha_t}{\alpha_t}$$

For fixed w we may define

$$\tilde{u}_t(x|y) = u_t^{\text{target}}(x) + w b_t \nabla \log p_t(y|x)$$

Rearranging yields

$$\tilde{u}_t(x|y) = (1-w) u_t^{\text{target}}(x) + w u_t^{\text{target}}(x|y)$$

This procedure is known as **classifier-free guidance**.

Classifier-Free Guidance Training

Observation: We may treat the unguided vector field as **conditioned on nothing.** But, **nothing is something:**

$$u_t^{\text{target}}(x) = u_t^{\text{target}}(x|y=\varnothing)$$

We may now train a single model $u_t^\theta(x|y),\ y \in \{\mathcal{Y}, \varnothing\}$ by re-using $\mathcal{L}_{\text{CFM}}^{\text{guided}}(\theta)$ and occasionally setting $y = \varnothing$:

$$\mathcal{L}_{\text{CFM}}^{\text{CFG}}(\theta) = \mathbb{E}_{\square}\|u_t^\theta(x|y) - u_t^{\text{target}}(x|z)\|^2$$

$$\square = (z,y) \sim p_{\text{data}}(z,y),\ \text{with prob. } \eta,\ y \leftarrow \varnothing,\ t \sim \text{Unif}[0,1),\ x \sim p_t(x|z)$$

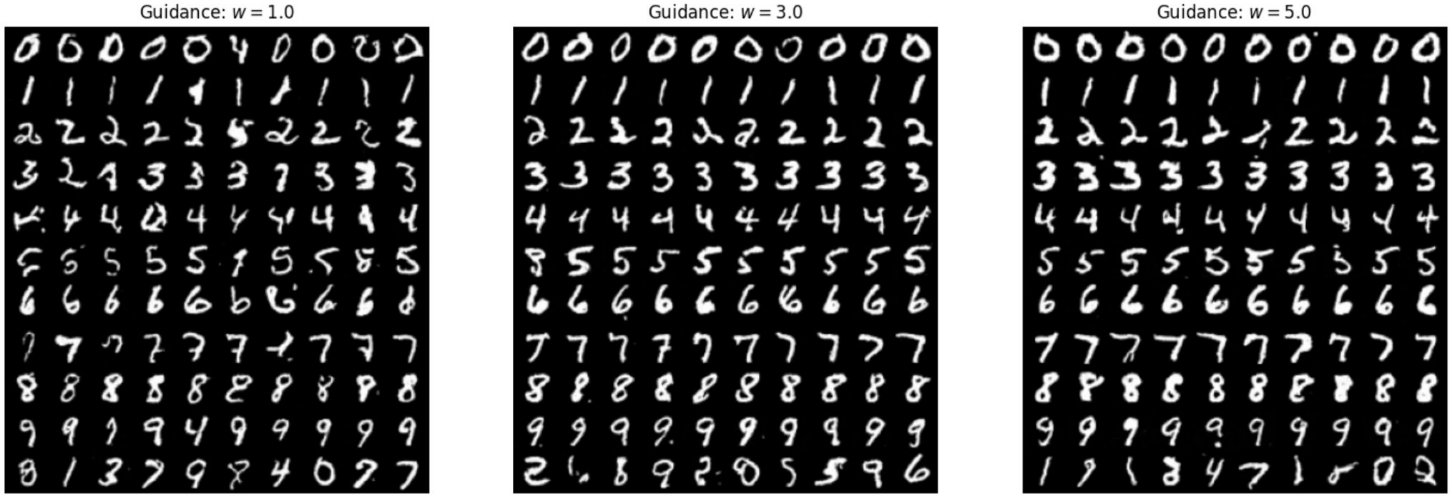


Image source: lab three!

Algorithm 8 Classifier-Free Guidance Sampling Procedure

Require: A trained guided vector field $u_t^\theta(x|y)$.

1: Select a prompt $y \in \mathcal{Y}$, or take $y = \varnothing$ for unguided sampling.

2: Select a **guidance scale** $w > 1$.

3: Initialize $X_0 \sim p_{\text{init}}$.

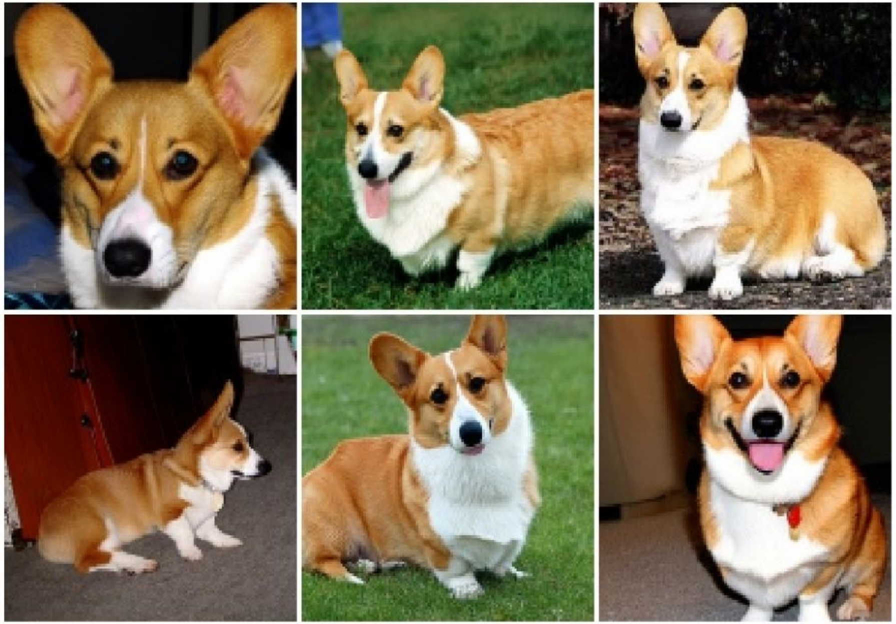
4: Simulate $\mathrm{d}X_t = \left[(1-w)u_t^\theta(X_t|\varnothing) + w u_t^\theta(X_t|y)\right]\mathrm{d}t$ from $t = 0$ to $t = 1$.

Image source: Classifier-free diffusion guidance [5].

Example: Classifier-Free Guidance

w=1.0

w=4.0



Part 2:
Architectural Considerations for Image Generation

Architectures for Image Generation

Recall: An image lives in $\mathbb{R}^{C_{\text{image}} \times H \times W}$

Question: An MLP is insufficient in such a high-dimensional space. What, then, should $u_t^\theta(x|y)$ look like?

Preview: We'll explore two choices: **U-Nets** (convolution based) and **diffusion transformers** (attention based).

Pay Attention: How is y **encoded**, **embedded**, and **processed**?

Architectures for Image Generation

Recall: An image lives in $\mathbb{R}^{C_{\text{image}} \times H \times W}$

Question: An MLP is insufficient in such a high-dimensional space. What, then, should $u_t^\theta(x|y)$ look like?

Preview: We'll explore two choices: **U-Nets** (convolution based) and **diffusion transformers** (attention based).

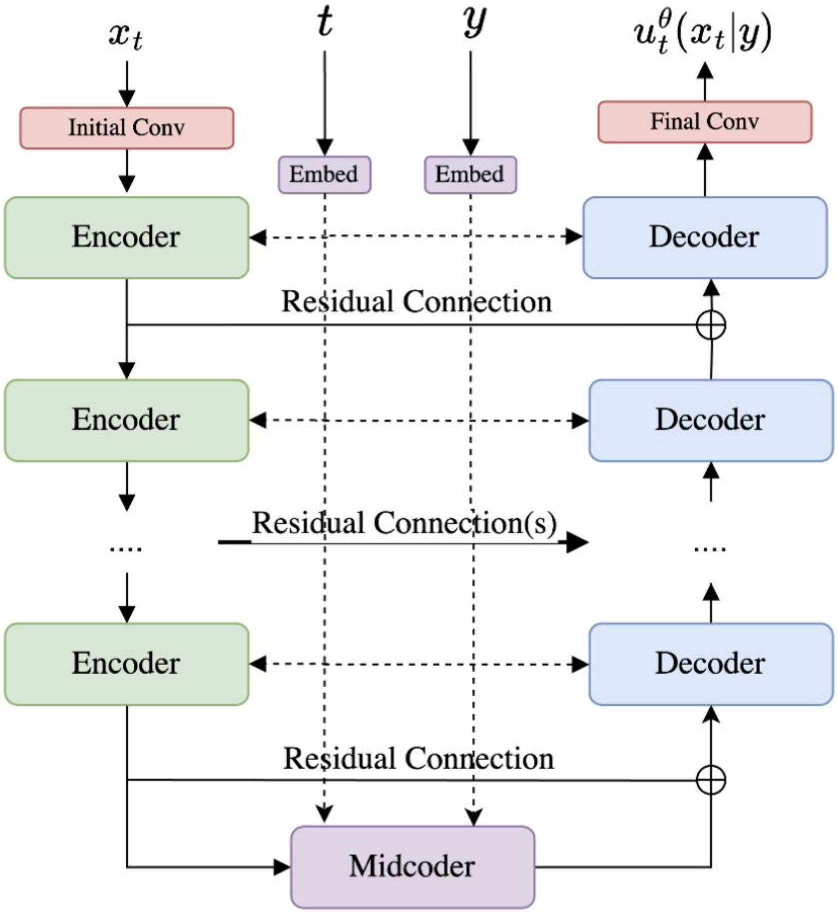
Pay Attention: How is y **encoded**, **embedded**, and **processed**?

Lab Three U-Net

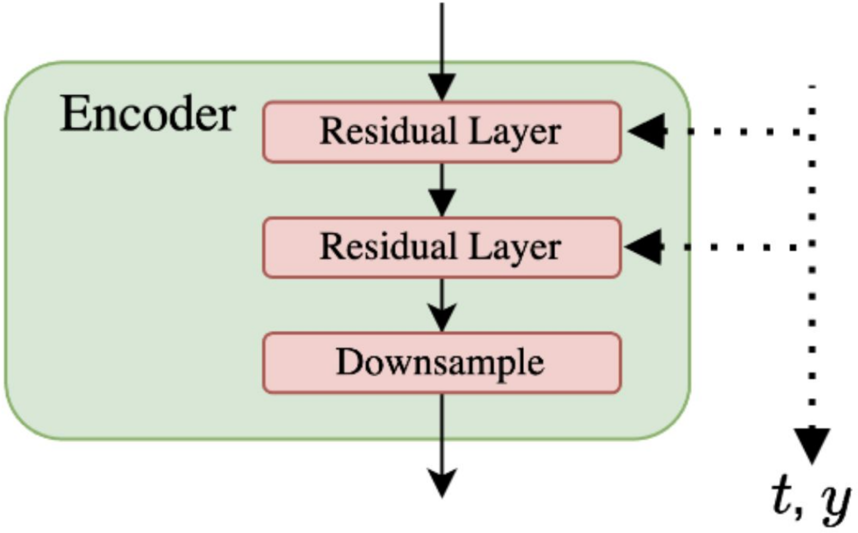
In lab three, we'll utilize the simplified **U-Net architecture** shown at right to build a generative model for the **MNIST dataset**.

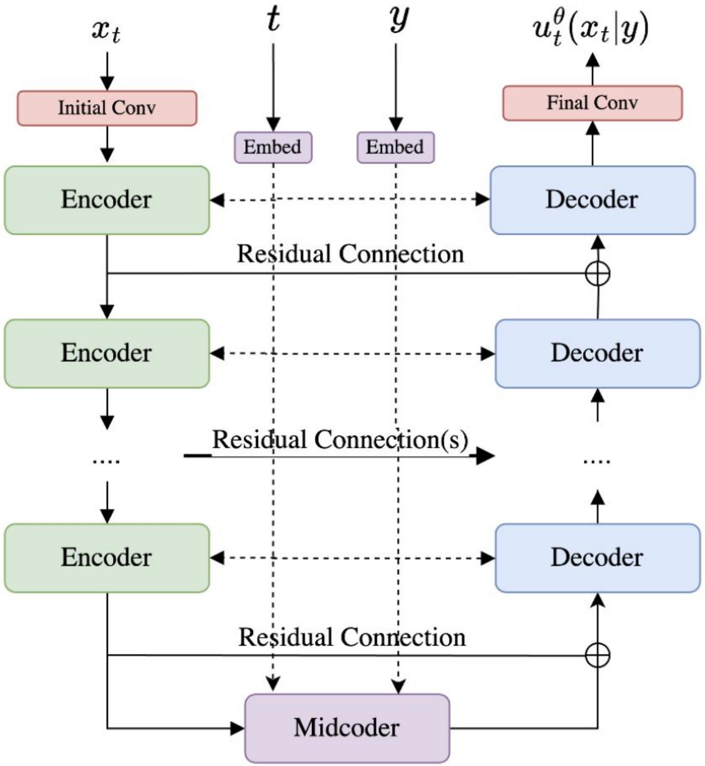
In this case $x_t \in \mathbb{R}^{1\times 32\times 32}$ and

$y \in \{0, 1, \ldots, 9, \varnothing\}$

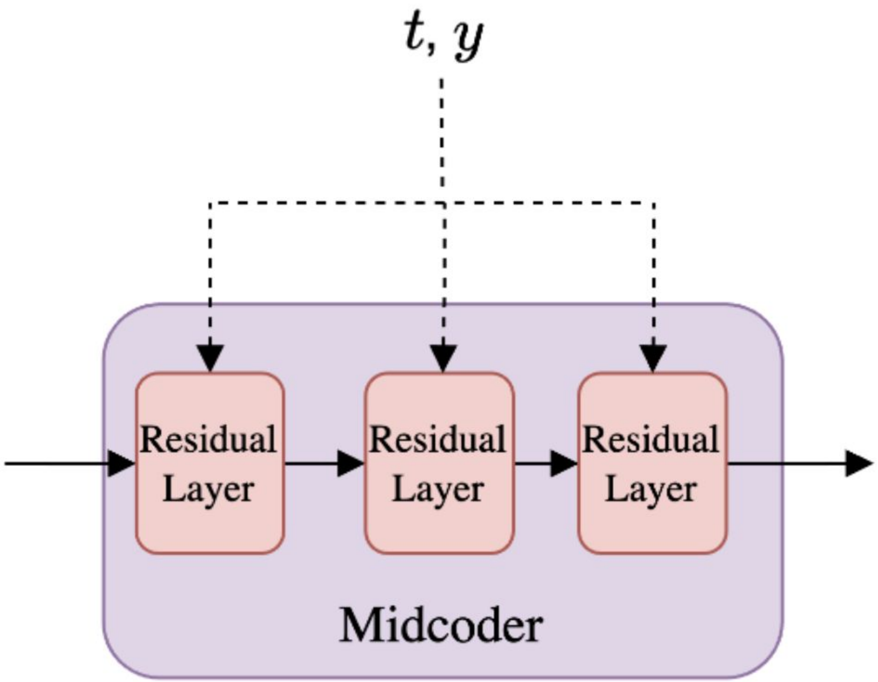


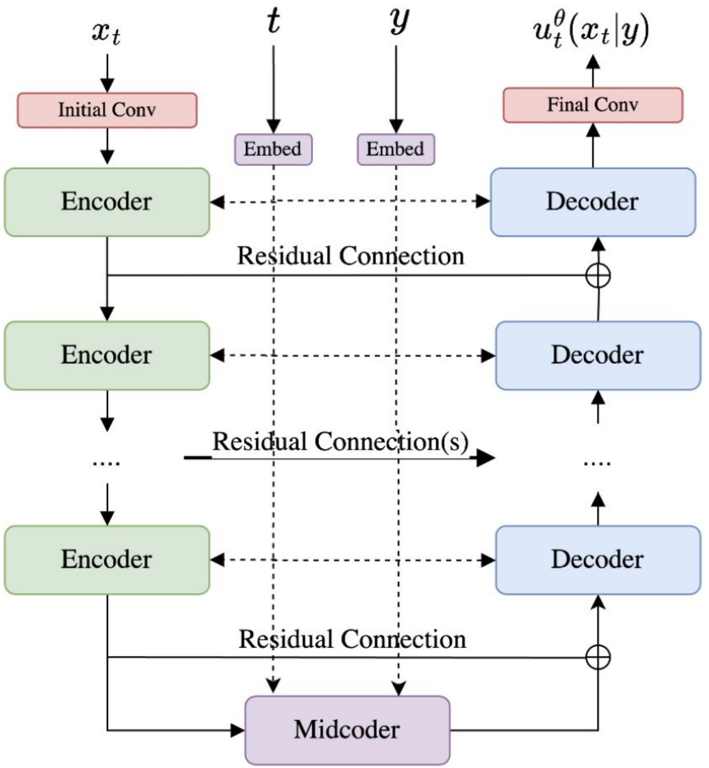
Lab Three U-Net: Encoder Layer



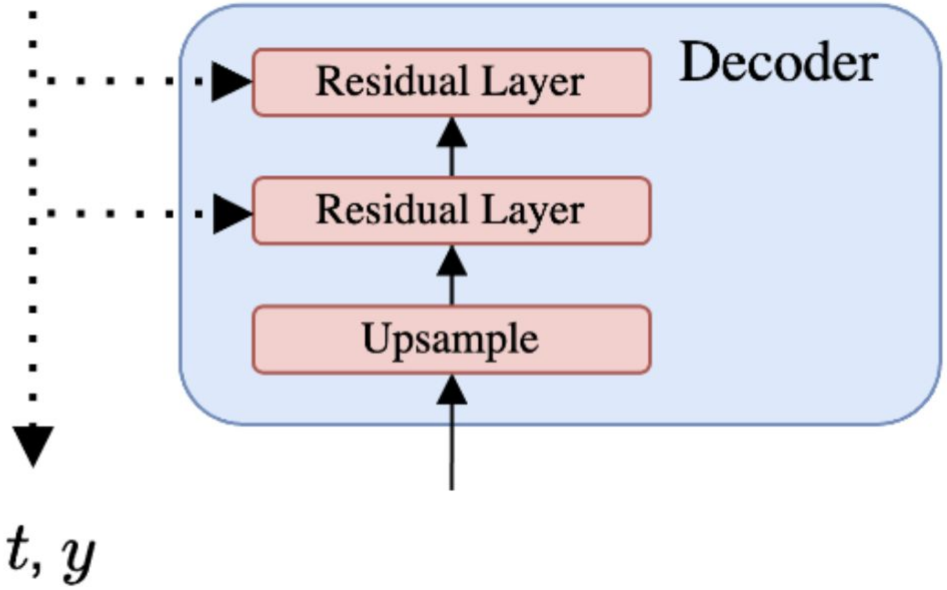


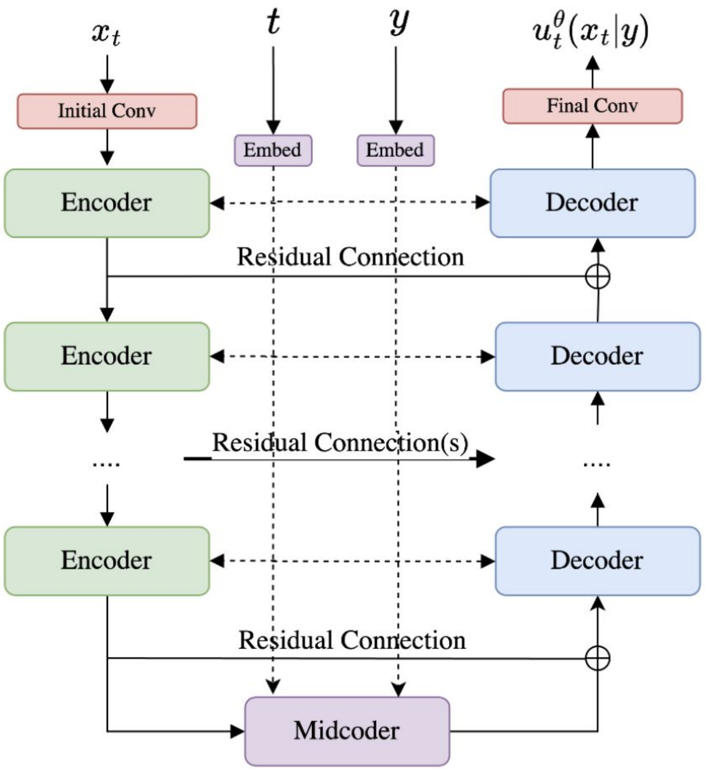
Lab Three U-Net: Midcoder Layer



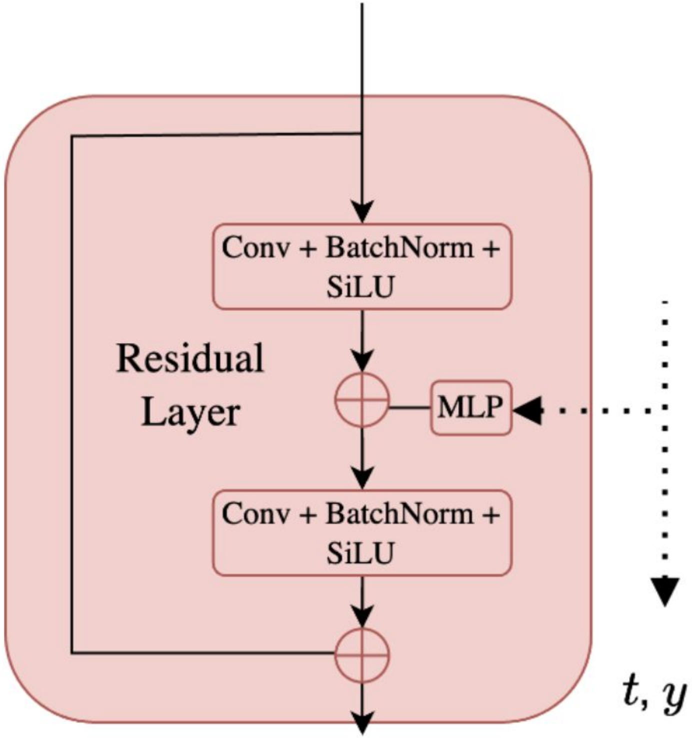


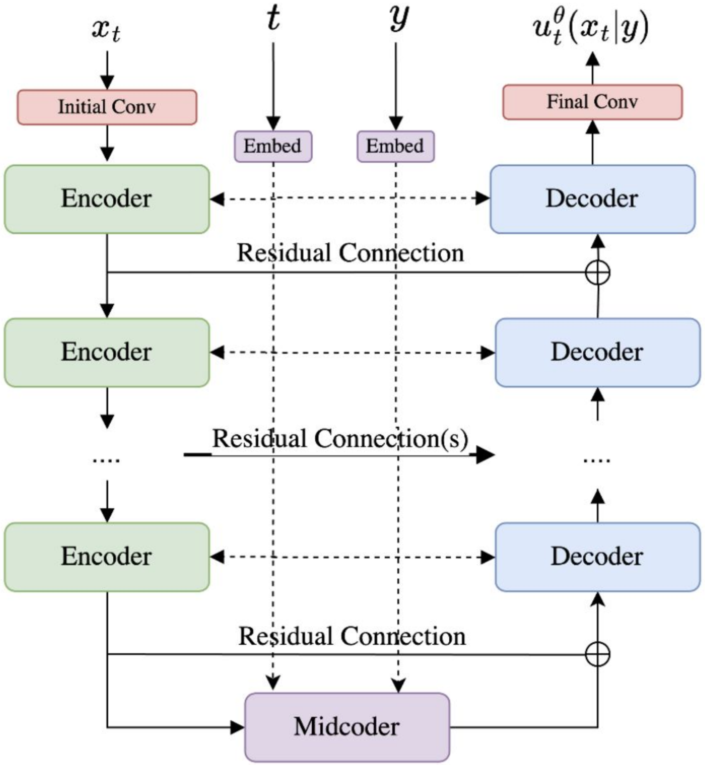
Lab Three U-Net: Decoder Layer





Lab Three U-Net: Residual Layer

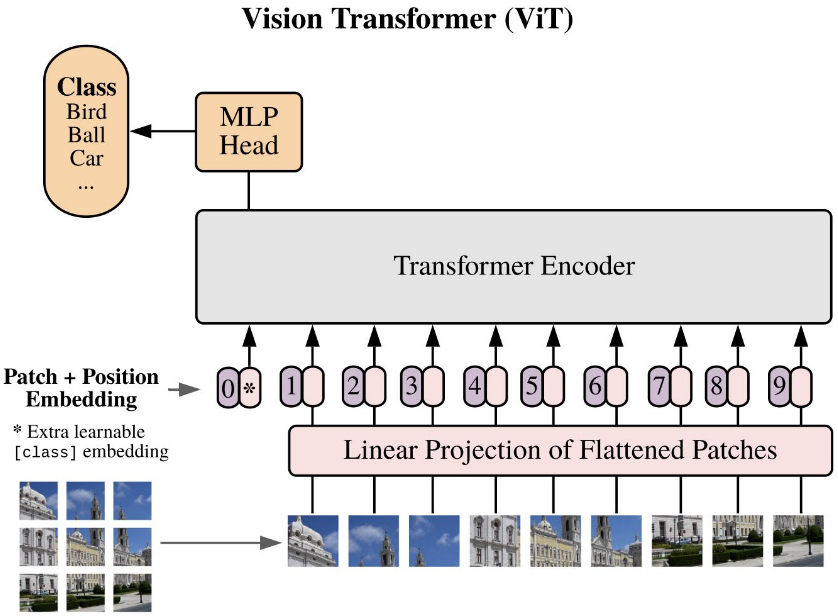


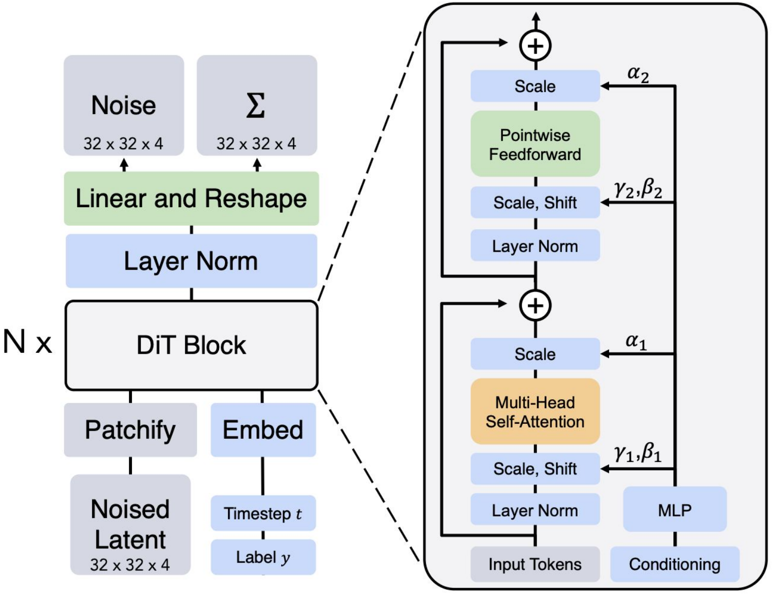


Diffusion Transformer (DiT)

Image sources: Vision transformer paper [2] (left), diffusion transformer paper [3] (right).

Idea: Divide an image into **patches** and **attend** between the patches. Based on the **vision transformer** (ViT).





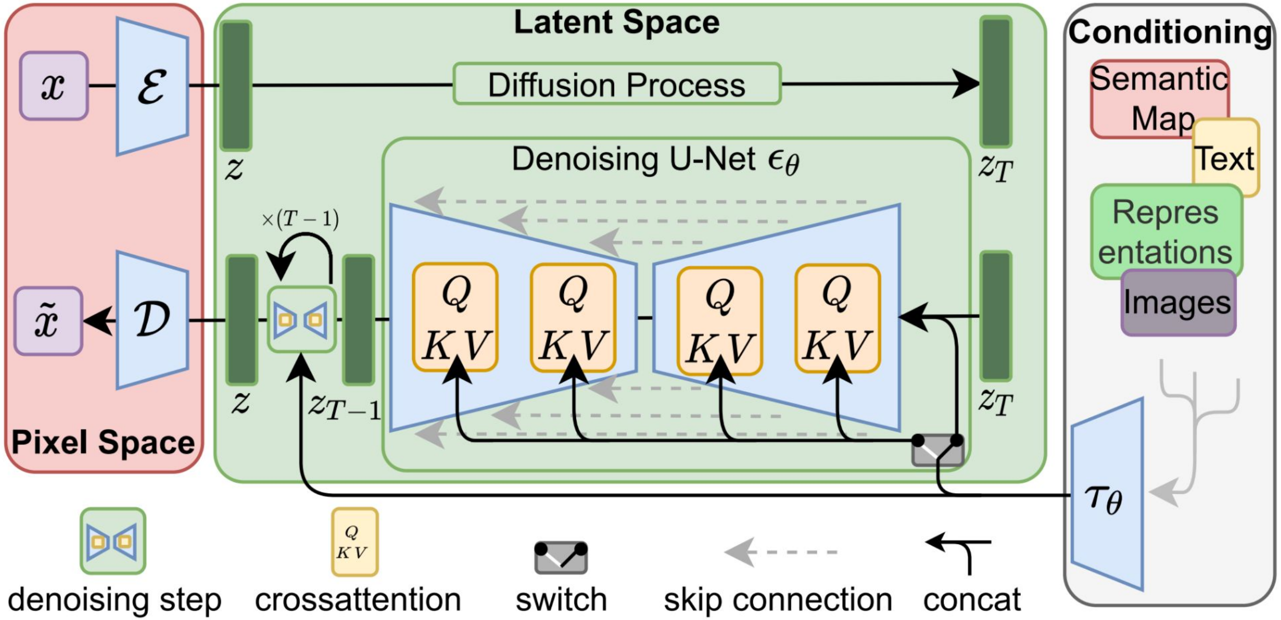
Latent Diffusion Transformer

DiT Block with adaLN-Zero

Image source: .High Resolution Image Synthesis with Latent Diffusion Models [4]

Generative Modeling in Latent Space

Idea: Train the generative model in the **latent space** of a pre-trained (variational) autoencoder.



Case Study: Stable Diffusion 3

Image source: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis [1]

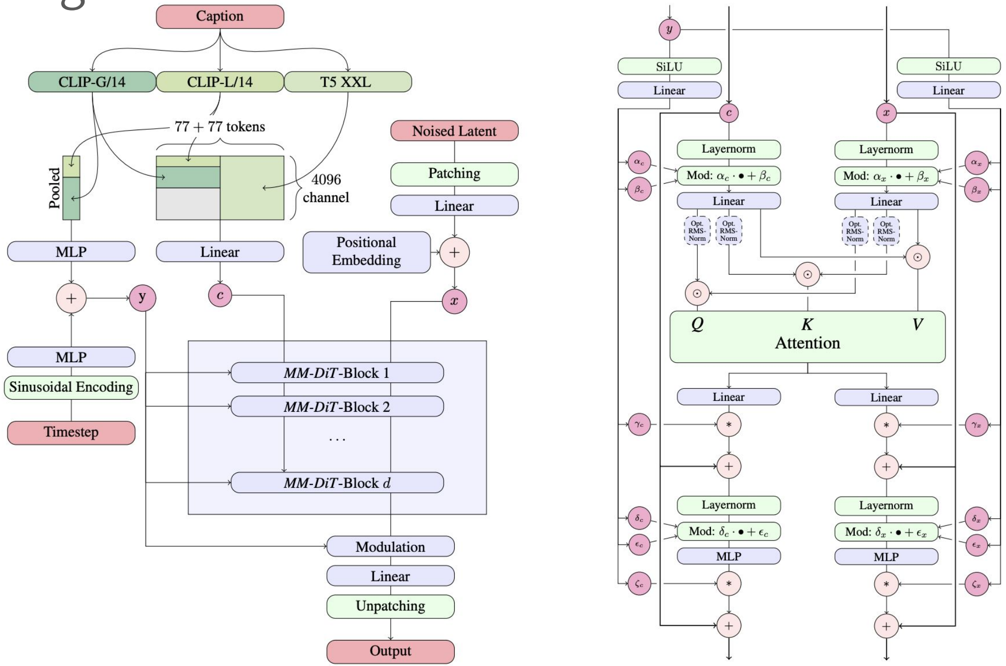
Ideas: Uses **pre-trained autoencoder**. Conditions on **CLIP** (coarse-grained) and **T5-XXL** (sequence-level) text embeddings via cross-attention. Extends DiT from class-conditioning to text-conditioning.

directly is intractable due to the marginalization in Equation 6, *Conditional Flow Matching* (see B.1),

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)}||v_\Theta(z,t) - u_t(z|\epsilon)||_2^2 \,, \quad (8)$$

with the conditional vector fields $u_t(z|\epsilon)$ provides an equivalent yet tractable objective.

Training objective used [1].



Next class:

Thursday (Jan 30), 11am-12:30pm

Robotics and Protein Design!

E25-111 (same room)

Office hours: Tuesday (37-212) & Wednesday (E25-111), 11am-12:30pm

References

1. **Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, https://arxiv.org/abs/2403.03206**
2. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, https://arxiv.org/abs/2010.11929**
3. **Scalable Diffusion Models with Transformers, https://arxiv.org/abs/2212.09748**
4. **High Resolution Image Synthesis with Latent Diffusion Models, https://arxiv.org/abs/2112.10752**
5. **Classifier-Free Diffusion Guidance, https://arxiv.org/abs/2207.12598**

Part 3:
Guest Talk!