Building an Image Generator.

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!

Conditional Generation: Cond on a prompt.

Guided.

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	???

Obs 1: Fix y $\Rightarrow$ Recover the unguided problem.

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

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

Obs 2: Now, let y vary.

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

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

Code: DataLoader.

Guided Sampling

(Prediction).

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.

Recall: A Gaussian cond prob path.

$$p_t(x|z) = \mathcal{N}(\alpha_t z; \beta_t^2 I_d). \qquad \alpha_1 = \beta_0 = 1. \quad \alpha_0 = \beta_1 = 0.$$

FACT:

$$u_t^{\text{target}}(x|y) = a_t x + b_t \nabla \log p_t(x|y).$$

$$(a_t, b_t) = \left(\frac{\dot{\alpha}_t}{\alpha_t},\ \frac{\alpha_t \beta_t^2 - \dot{\beta}_t \beta_t \alpha_t}{\alpha_t} \right).$$

Check previous notes.

Recall: Bayes Rule. $\nabla \log p_t(x|y) = \nabla \log \frac{p_t(x)\, p_t(y|x)}{p_t(y)}$

$= \nabla \log p_t(x)\, p_t(y|x) - \underbrace{\nabla \log p_t(y)}_{=\,0}$.

Plug in Bayes Rule:

$$u_t^{\text{target}}(x|y) = a_t x + b_t \left(\nabla \log p_t(x) + \nabla \log p_t(y|x) \right)$$

$$= \underbrace{\left(a_t x + b_t \nabla \log p_t(x)\right)}_{u_t^{\text{target}}(x)} + b_t \nabla \log p_t(y|x).$$

↑ guidance with classifier.

$$= u_t^{\text{target}}(x) + b_t \nabla \log p_t(y|x).$$

⇓

The term is a sort of "classifier".

All signals of y.

To improve the significance of y:

Choose a "guidance scale" $w > 1$.

$$v_t(x|y) = u_t^{\text{target}}(x) + \boxed{w}\, b_t \nabla \log p_t(y|x).$$

$$= u_t^{\text{target}}(x) + w b_t \left(-\nabla \log p_t(x) + \nabla \log p_t(x|y) \right)$$ ↩ Plug back Bayes Rule.

$$= u_t^{\text{target}}(x) - w a_t x + w a_t x + w b_t \left(\nabla \log p_t(x|y) - \nabla \log p_t(x) \right)$$

$$= u_t^{\text{target}}(x) - w\left(a_t x + b_t \nabla \log p_t(x)\right) + w\left(a_t x + b_t \nabla \log p_t(x|y)\right)$$

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

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 $y = \varnothing$:

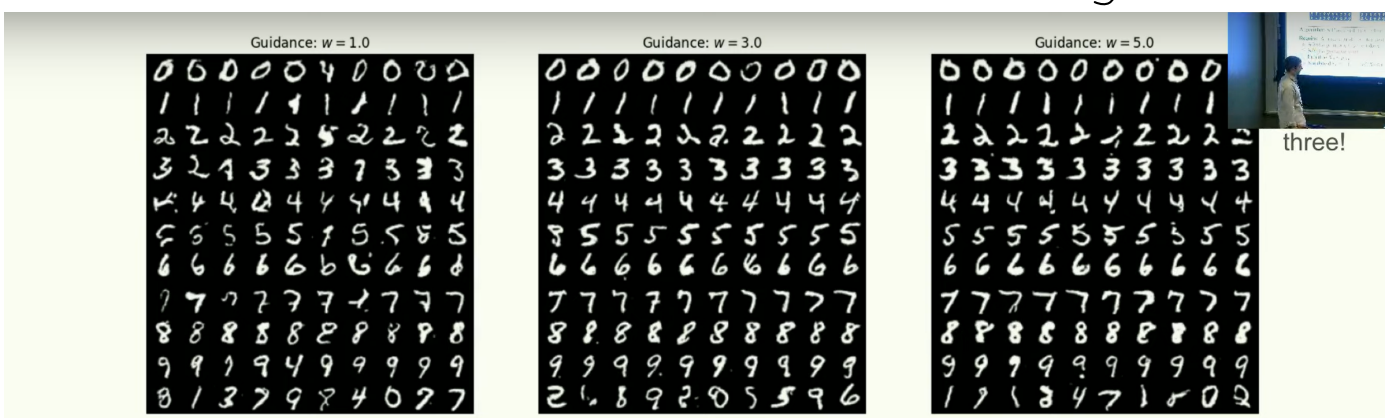
只需 train 一个 model.

在 train 过程中随机插入 non-prompt.

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

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

$Pr[y=\phi] = \eta$. Randomly insert.



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 $t = 1$.

在 Prediction 时引入 $y = \phi$.

$w \uparrow \Rightarrow$ Clear generation.

tradeoff between { quality / diversity } of generation.

Architecture. (for $u_t^\theta(x|y)$)

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).

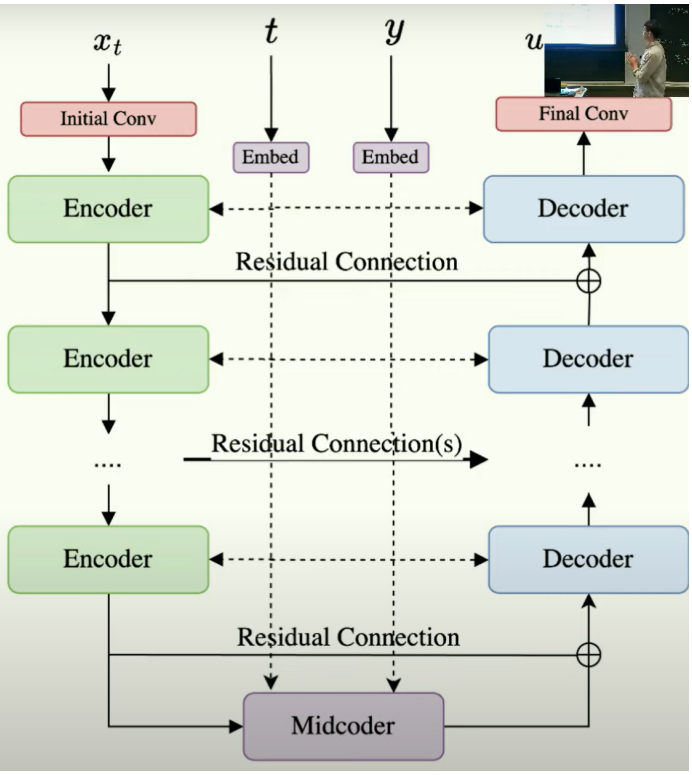
Pay Attention: How is y **encoded**, **embedded**, and **processed**?

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

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



paper [3] (right).

Diffusion Transformer (DiT)

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