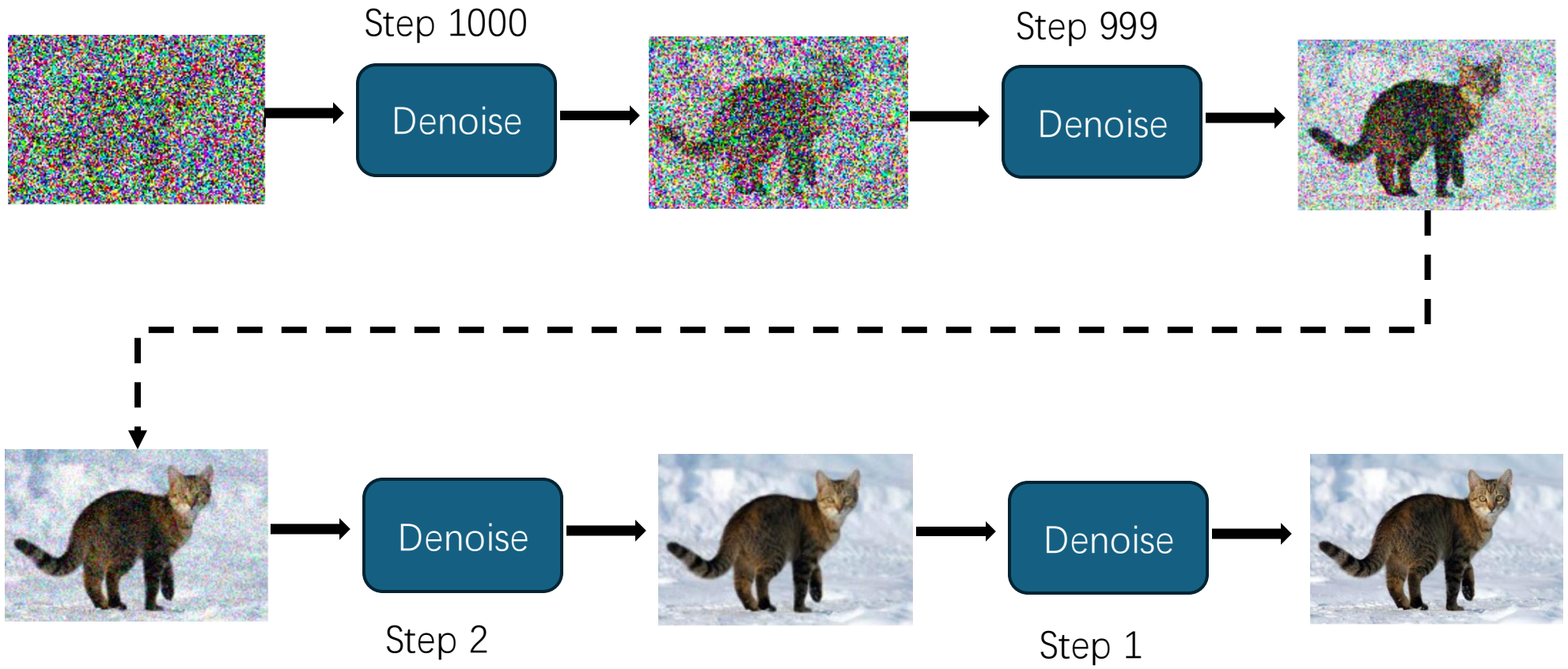
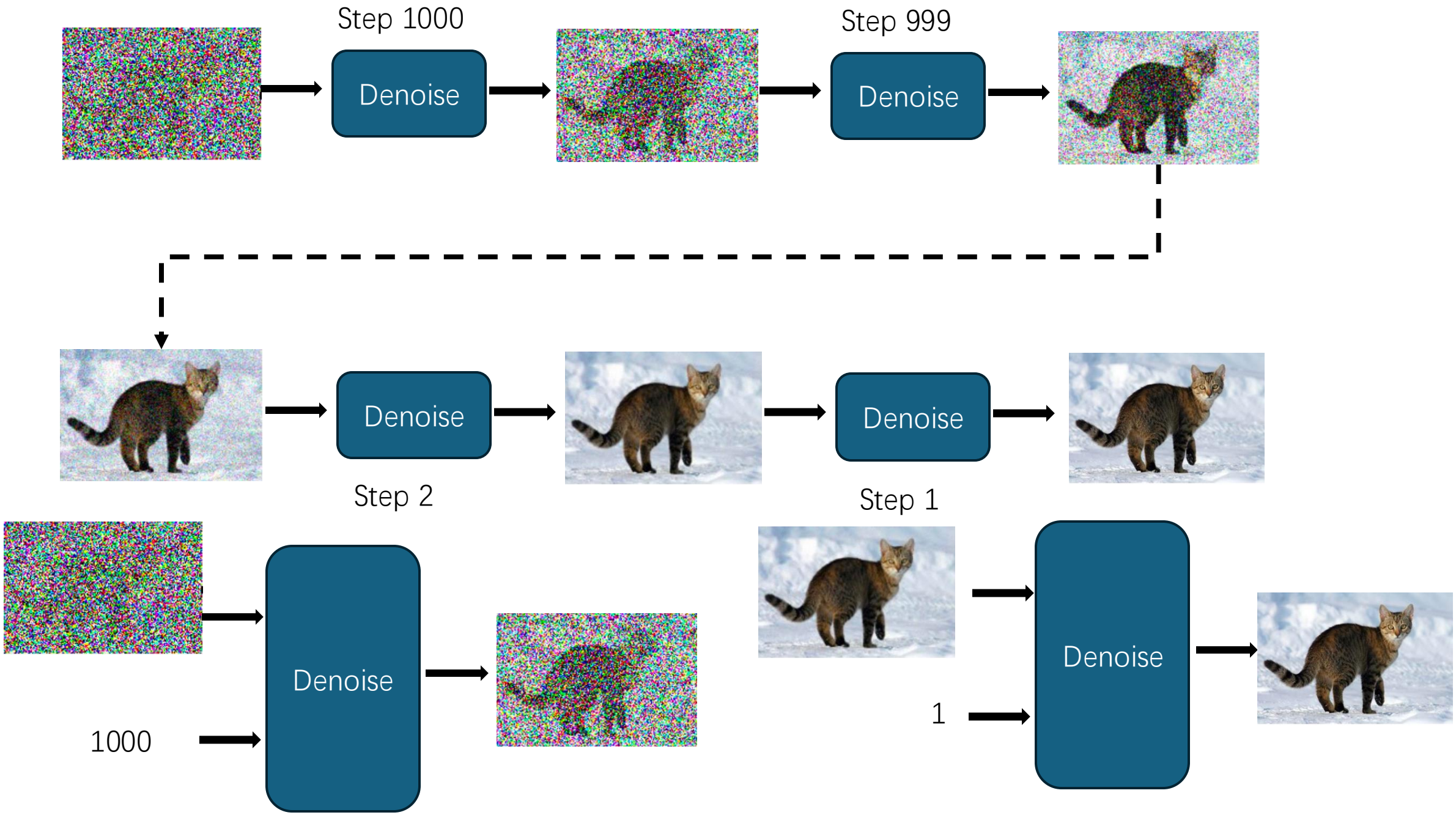
Diffusion Model

Shared by jerryrzeng

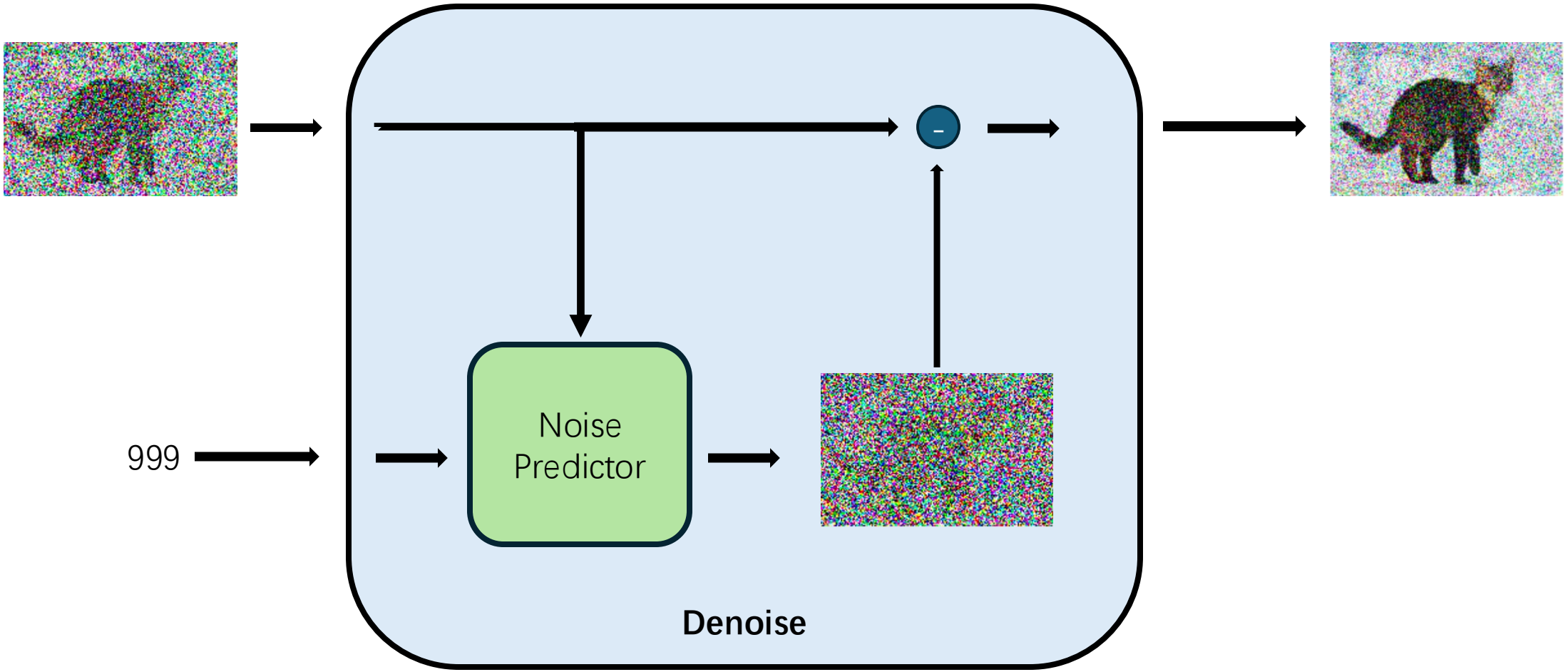
Denoising Diffusion Probabilistic Models (DDPM)
[2006.11239] Denoising Diffusion Probabilistic Models (arxiv.org)

Diffusion Model是如何运作的?

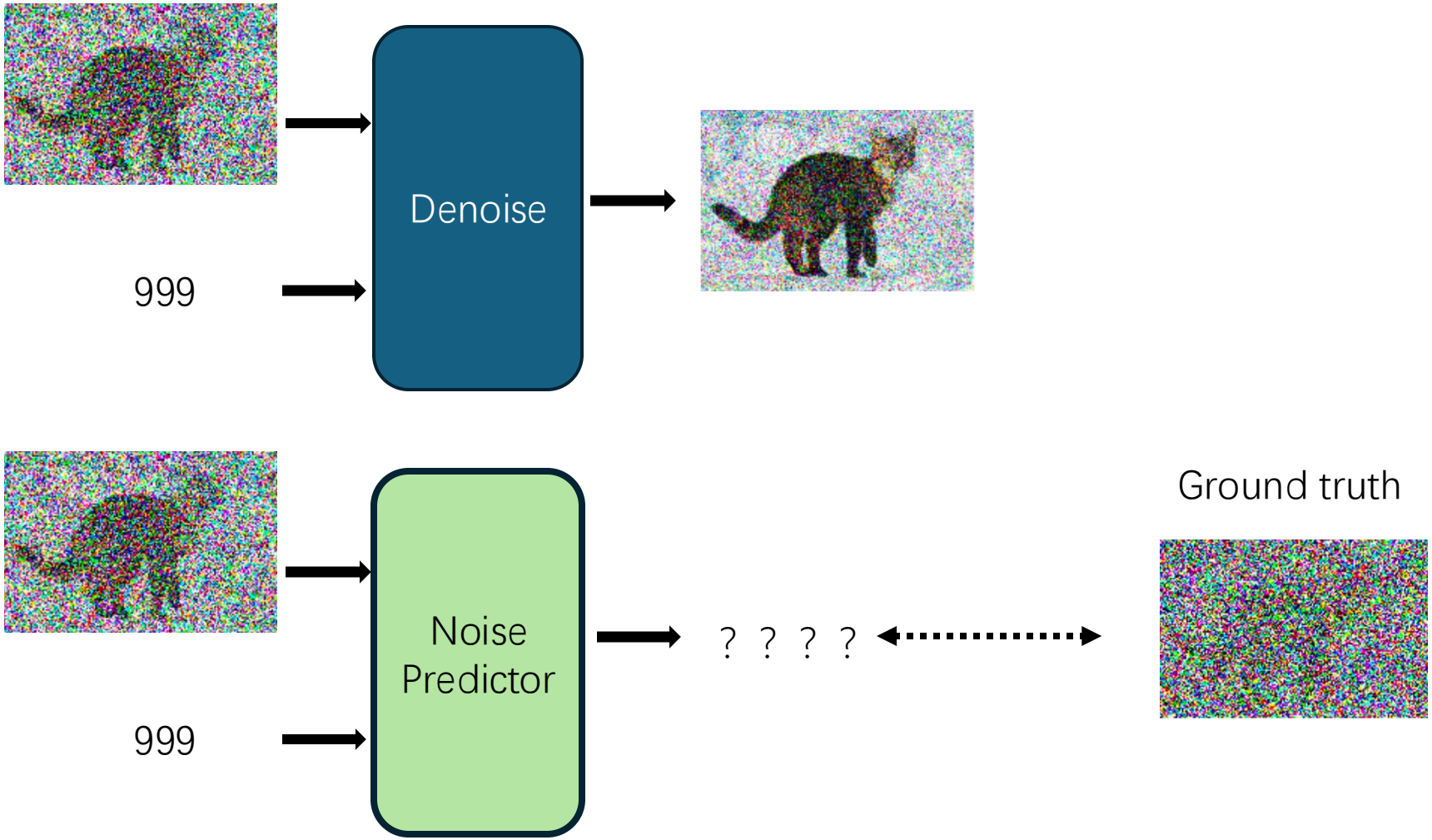



Step 1000
Denoise
Step 999
Denoise
Denoise
Denoise
Step 2
Step 1
1000
Denoise
1
Denoise

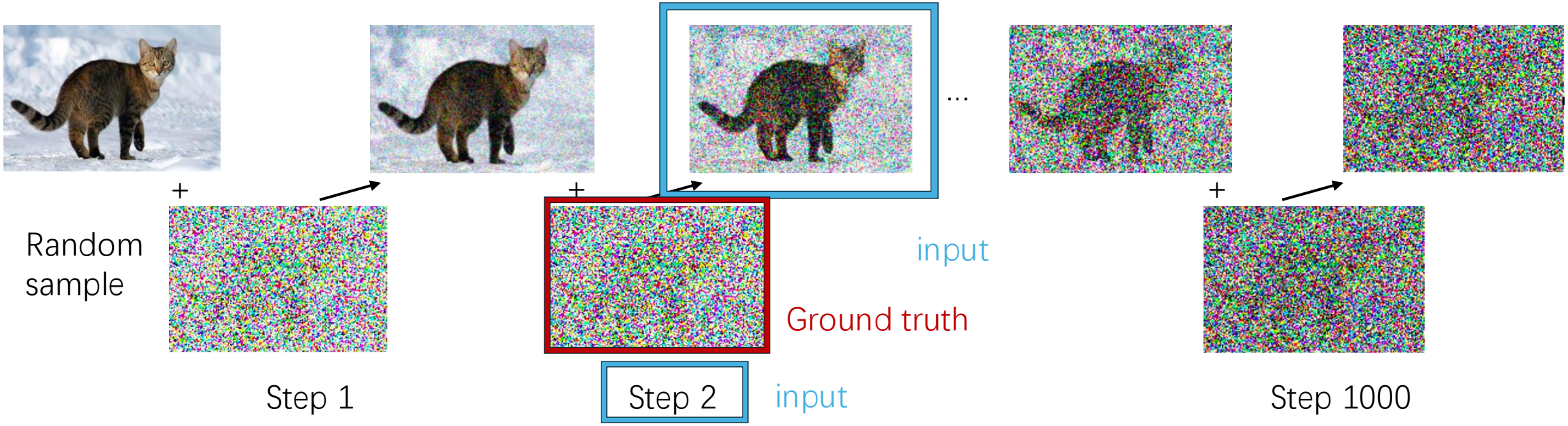
Denoise 模块内部实际做的事情

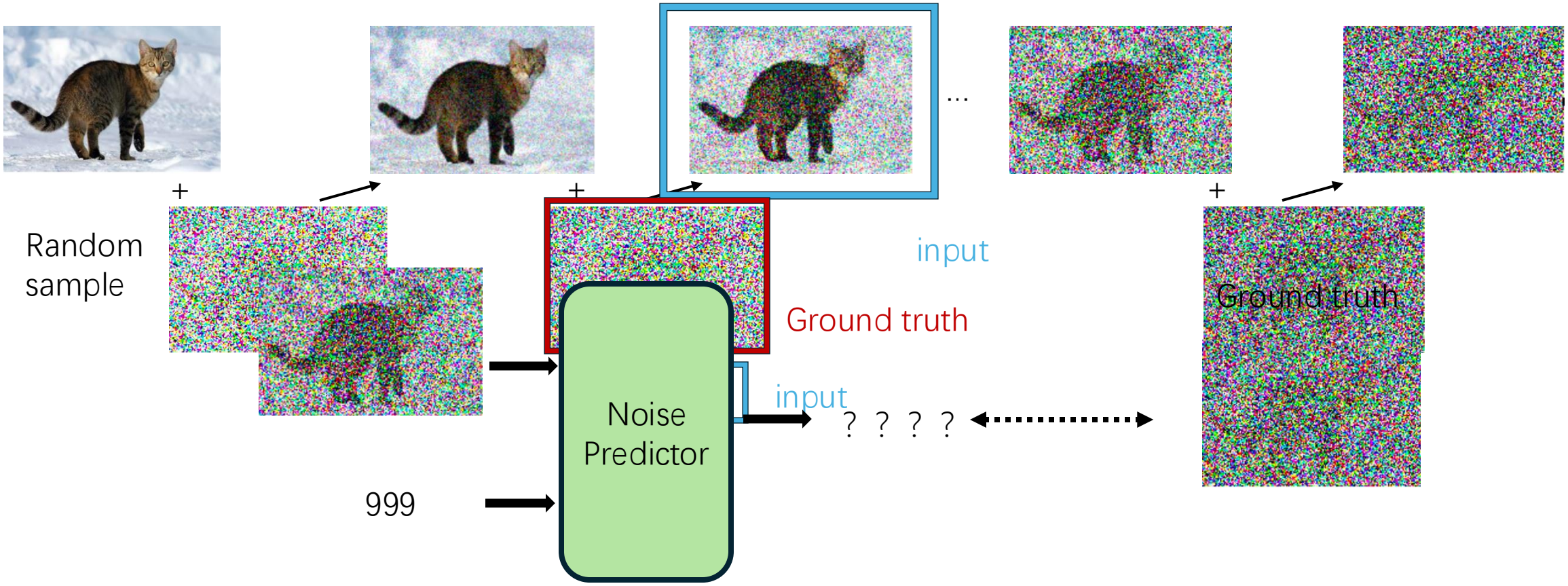


如何训练Noise Predictor



Forward Process (Diffusion Process)



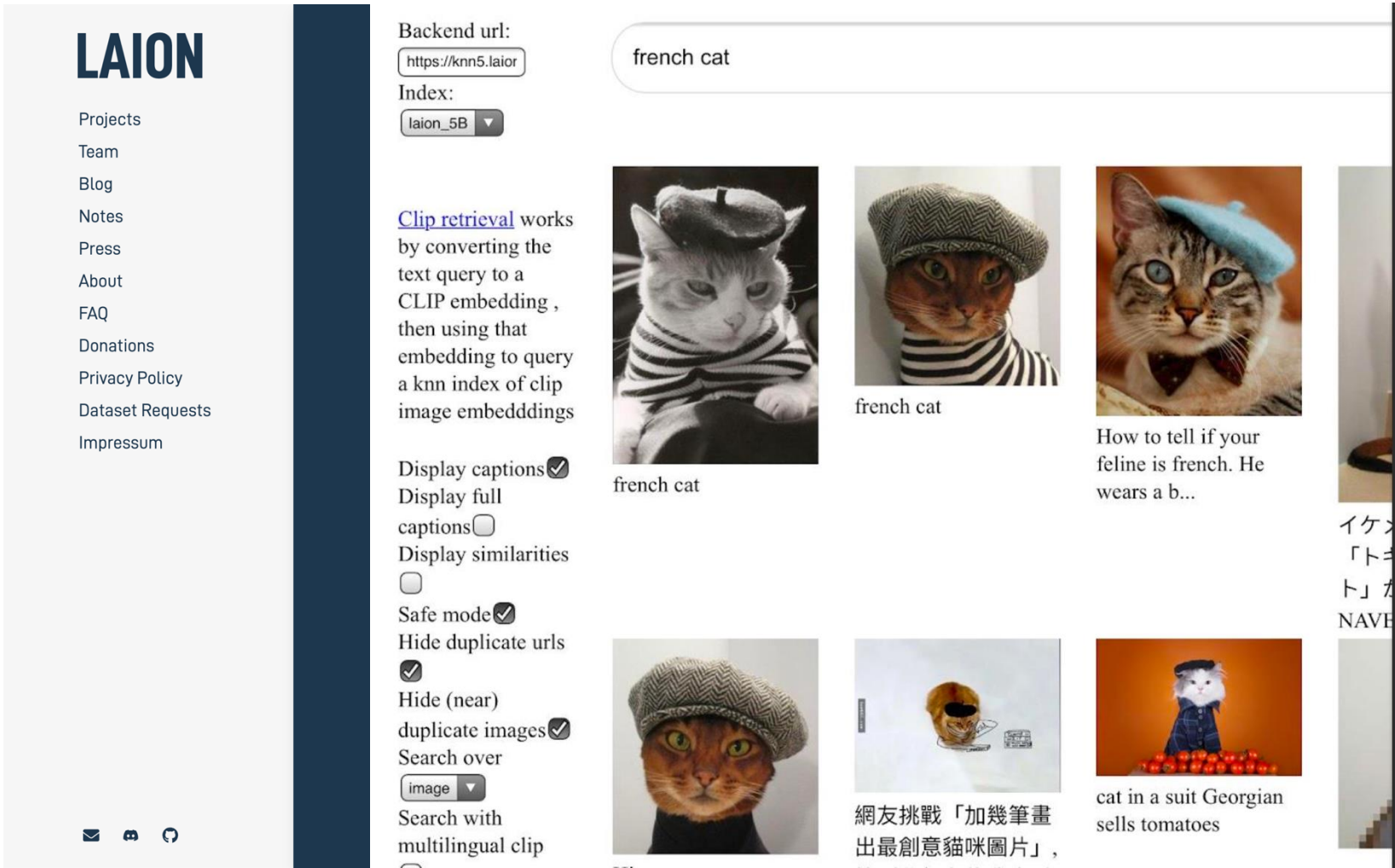

Random sample
+
+
+
...
input
Ground truth
input
? ? ? ?
Noise Predictor
999
Ground truth

Text-to-Image

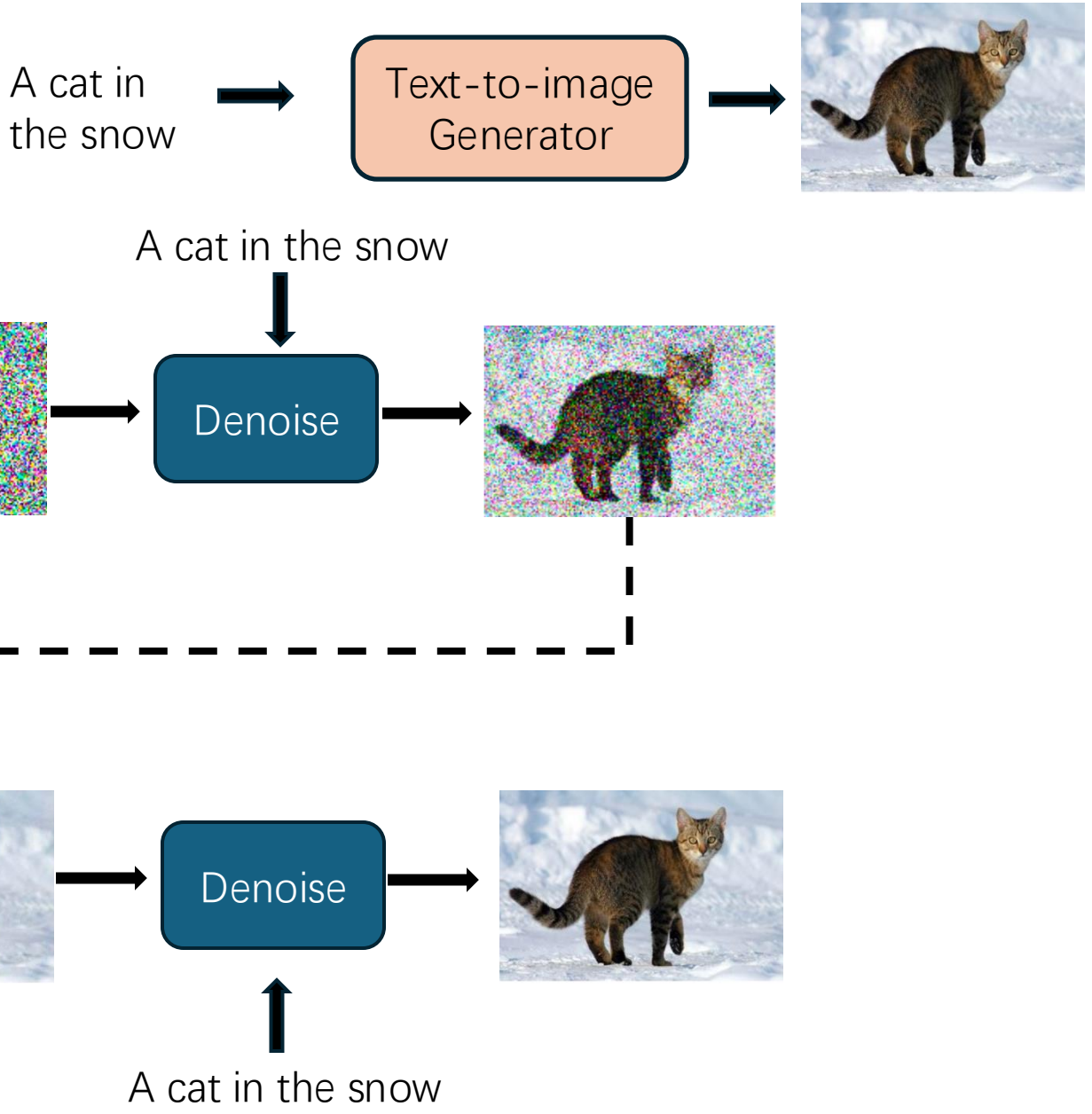
[LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS | LAION]

A cat in the snow → Text-to-image Generator →

ImageNet	LAION
1M	5.85B

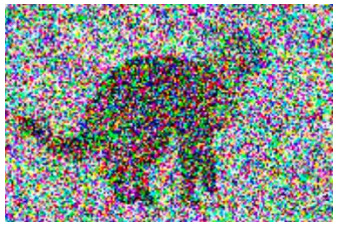


Text-to-Image



Text-to-Image

A cat in the snow → Text-to-image Generator → 

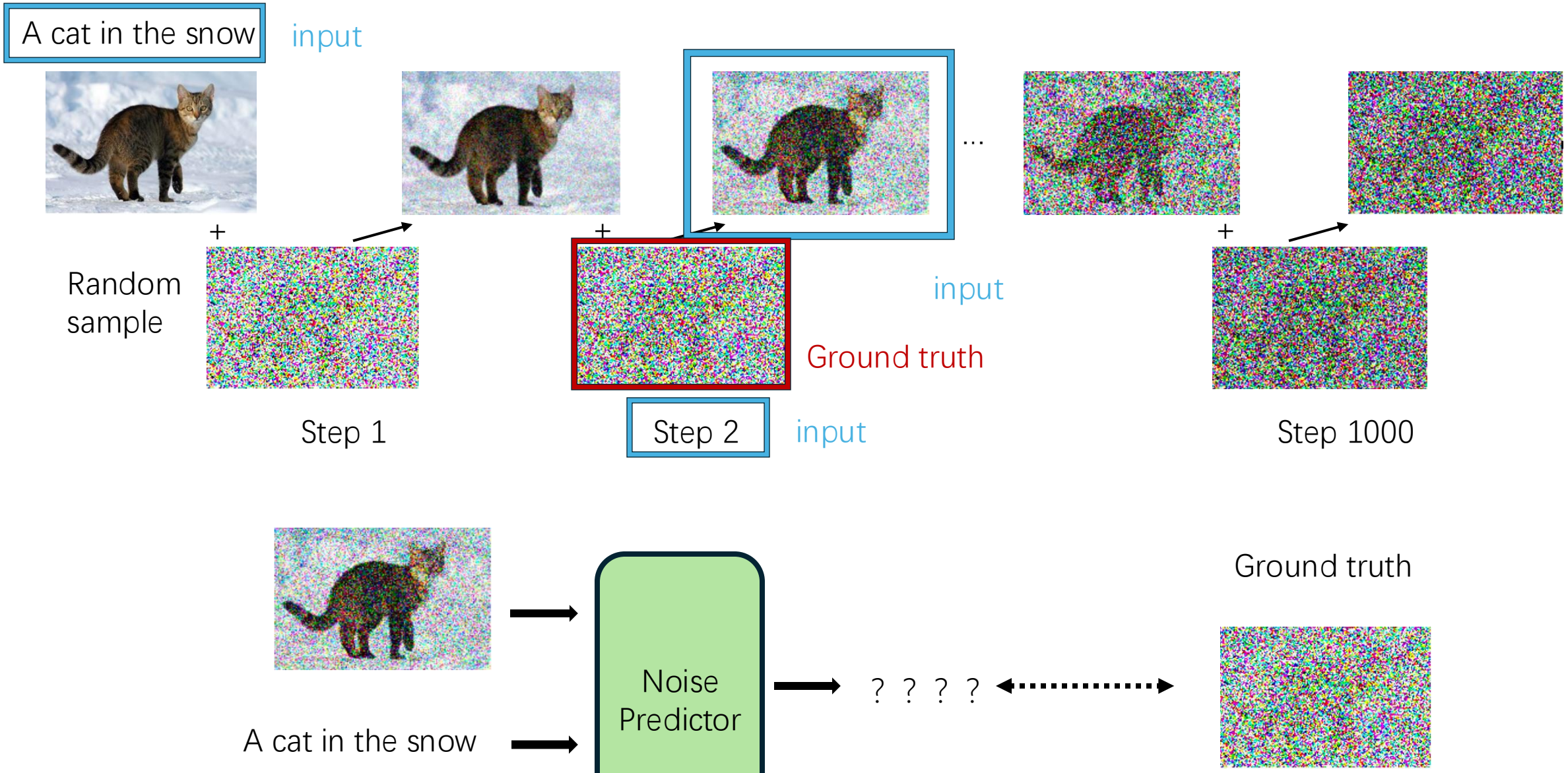


A cat in the snow

999

Noise Predictor

Denoise


A cat in the snow
input
Random sample
+
Step 1
+
input
Ground truth
Step 2
input
...
+
Step 1000
A cat in the snow
2
Noise Predictor
? ? ? ?
Ground truth

Algorithm 1 Training

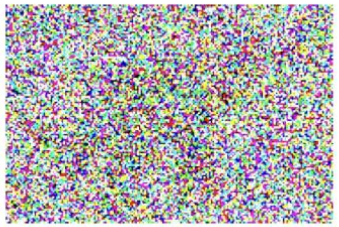
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Training

x_0: clean image



ε: noise

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ ← Sample clean image
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ← Sample a noise
5: Take gradient descent step on

$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right) \right\|^2$$

6: **until** converged

Noise image

Target Noise

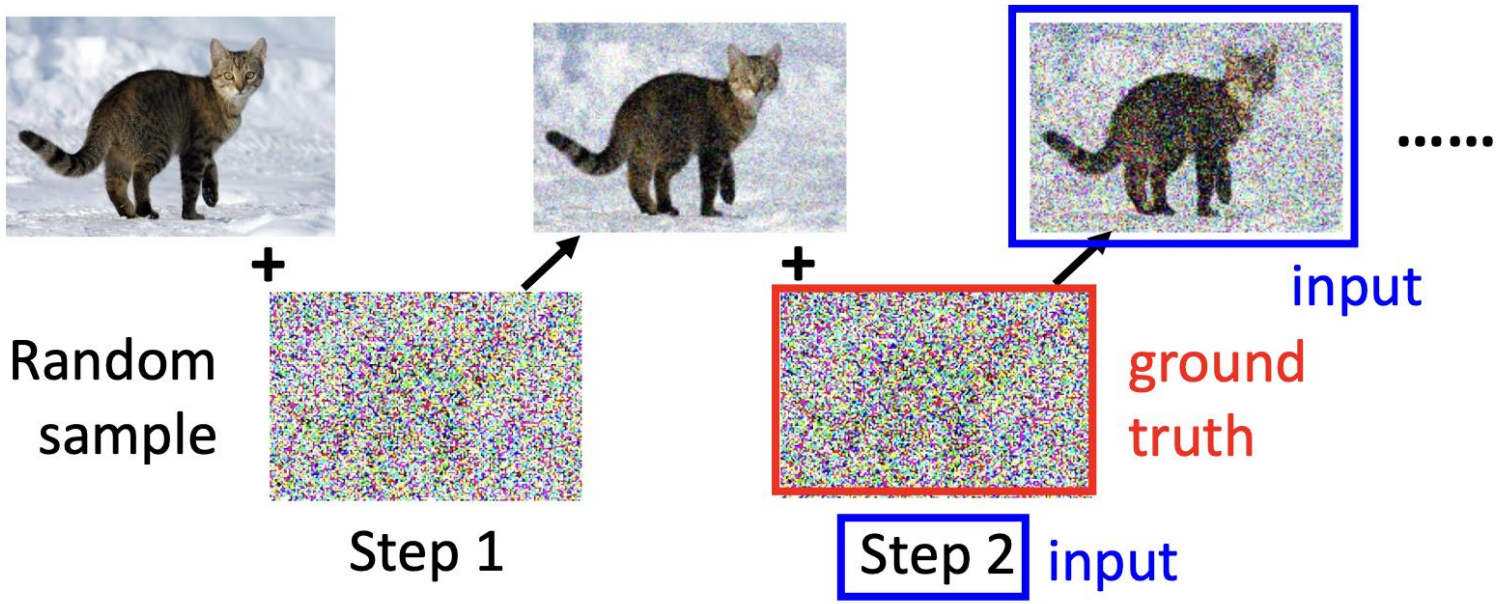
Noise Predictor

$\bar{\alpha}_1, \bar{\alpha}_2, \dots \bar{\alpha}_T$

smaller

Training

理论上



实际中

$$\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon = \text{input}$$

x_0

ε

ground
truth

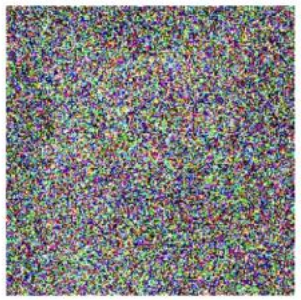
Inference

Algorithm 2 Sampling

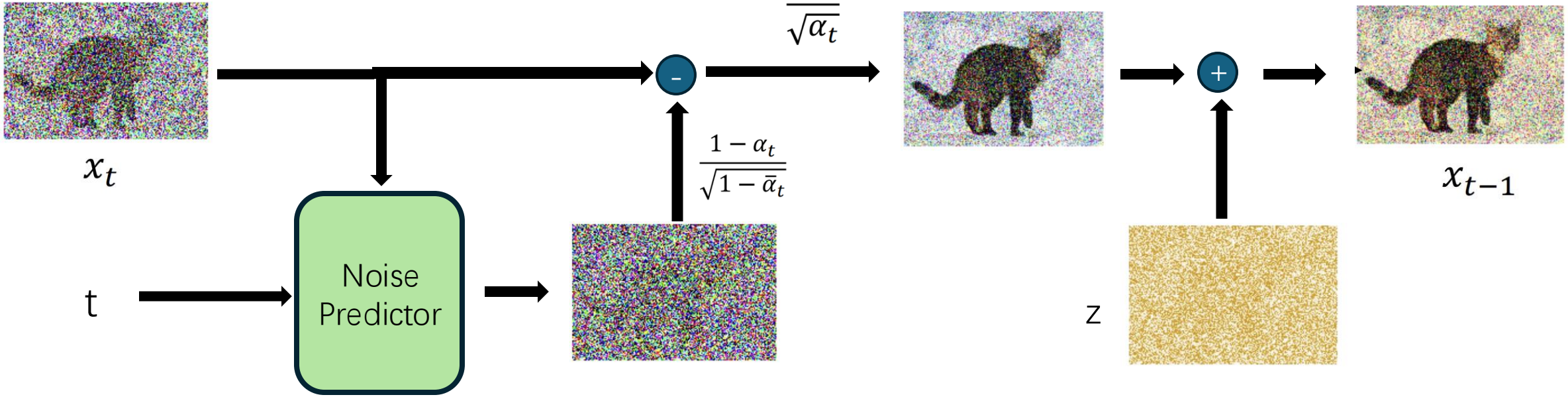
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

$\bar{\alpha}_1, \bar{\alpha}_2, \ldots \bar{\alpha}_T$

$\alpha_1, \alpha_2, \ldots \alpha_T$



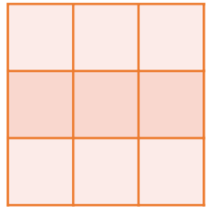
x_T

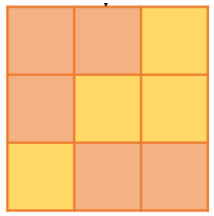


Stable Diffusion

A cat in the snow → Text-to-image Generator → 

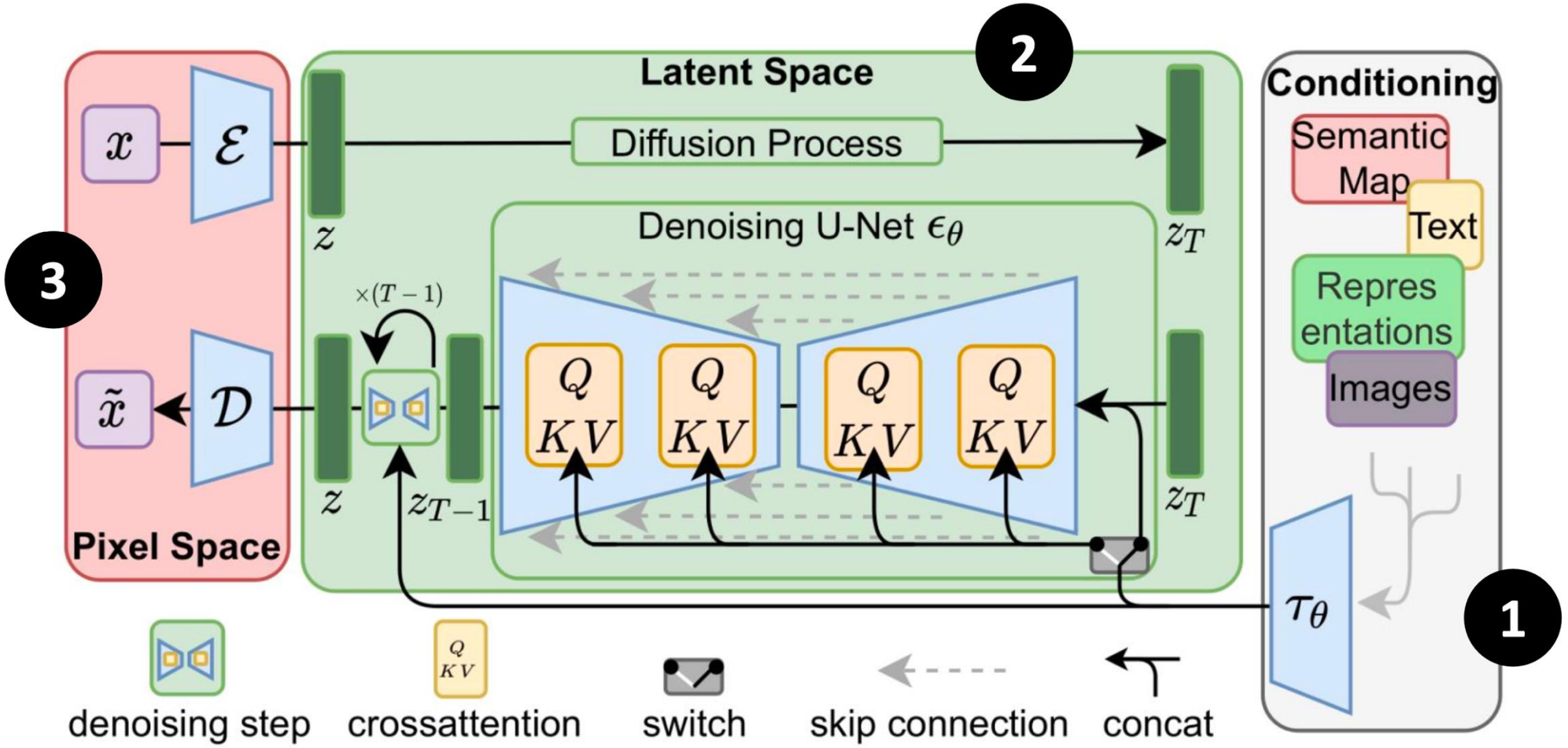
A cat in the snow → 1 Text Encoder → 

 → Generation Model 2

中间产物
图片的压缩版本  → Decoder 3 → 

Stable Diffusion

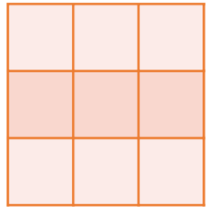
[2112.10752] High-Resolution Image Synthesis with Latent Diffusion Models (arxiv.org)

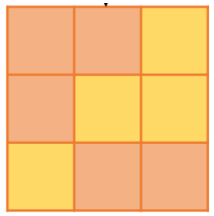


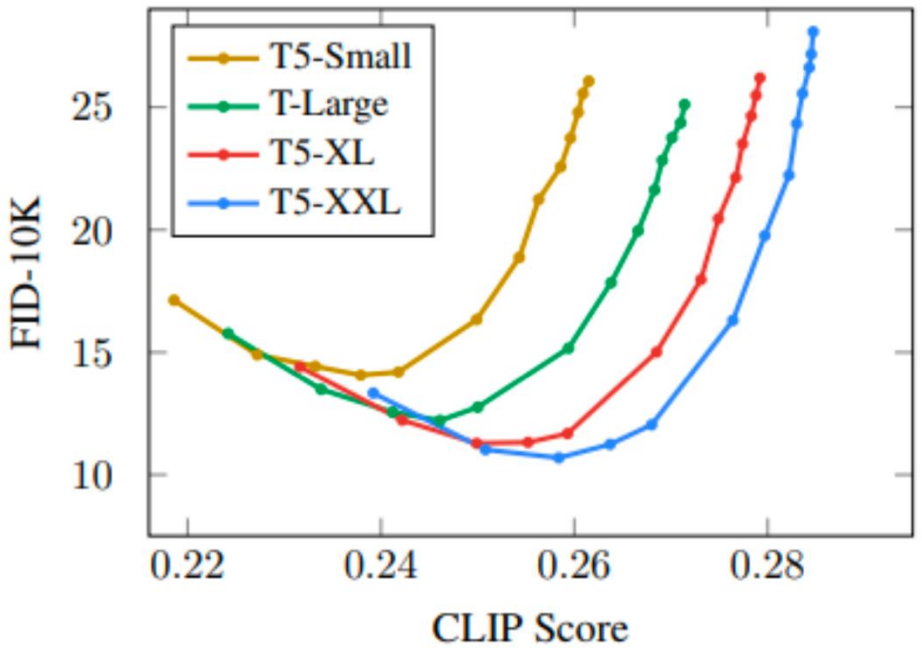
Framework

A cat in the snow → Text-to-image Generator → 

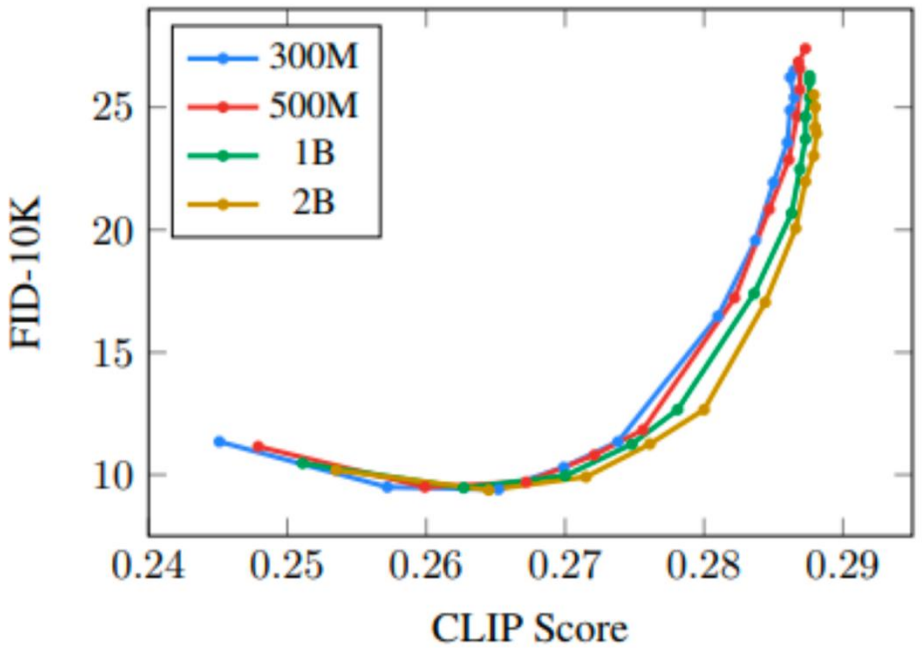
A cat in the snow → 1 Text Encoder → 

 → Generation Model 2

中间产物
图片的压缩版本  → Decoder 3 → 



(a) Impact of encoder size.



(b) Impact of U-Net size.

[2205.11487] Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (arxiv.org)

Fréchet Inception Distance (FID)

[1706.08500] GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium (arxiv.org)

CNN

softmax

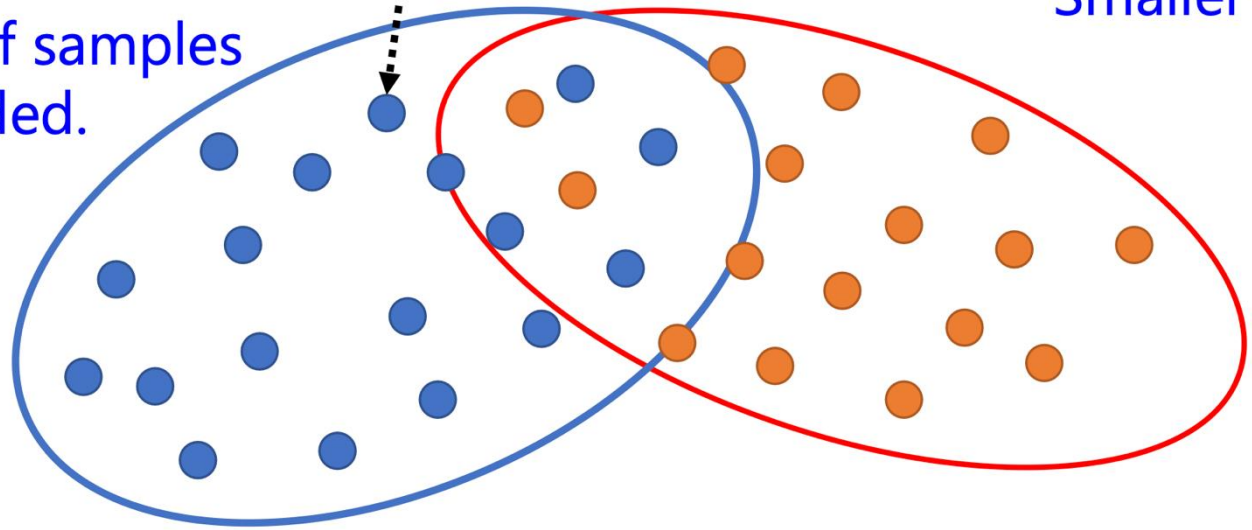
red points: real images

blue points: generated images

FID = Fréchet distance between the two **Gaussians** ???

Smaller is better

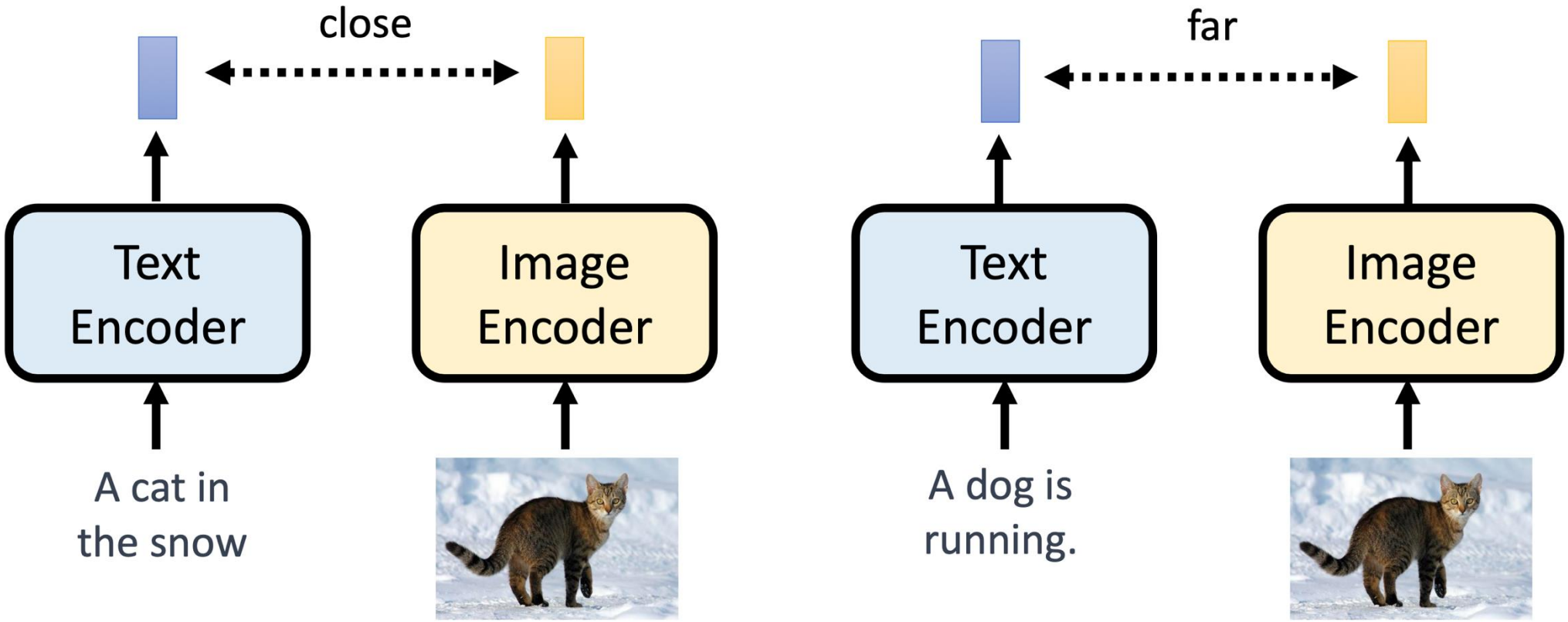
A lot of samples is needed.



Contrastive Language-Image Pre-Training (CLIP)

[2103.00020] Learning Transferable Visual Models From Natural Language Supervision (arxiv.org)

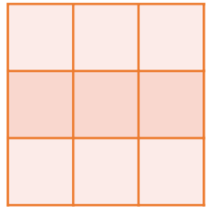
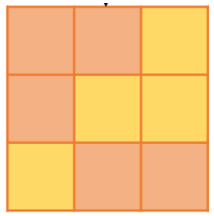
400 million image-text pairs



Framework

A cat in the snow → Text-to-image Generator → 

A cat in the snow → 1 Text Encoder → 

 → Generation Model 2 → 

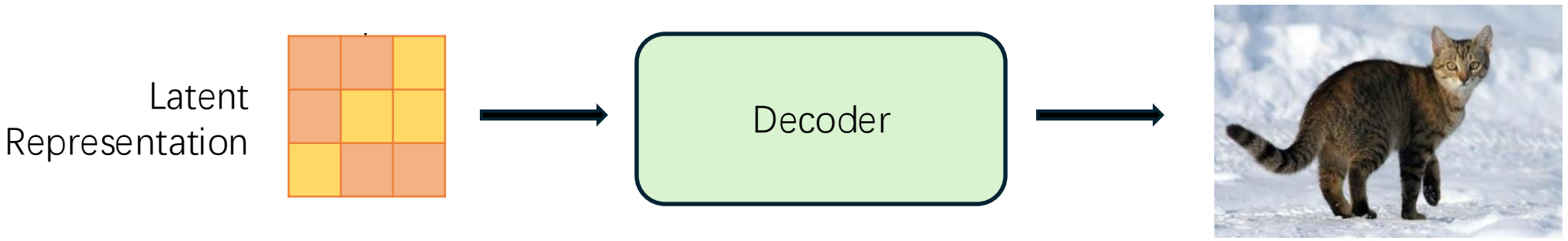
Decoder 3

Decoder can be trained without labelled data

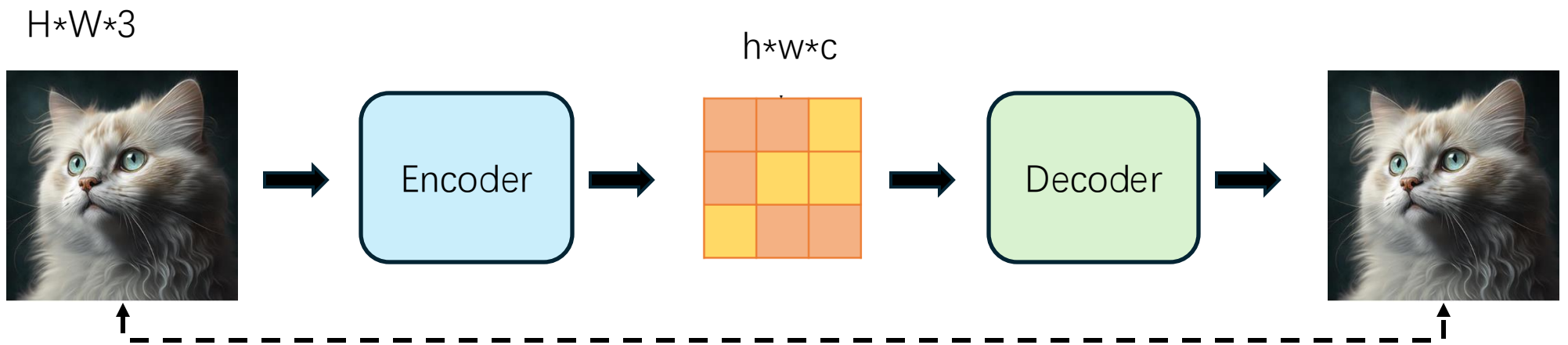
中间产物为小图

Decoder

中间产物为 Latent Representation



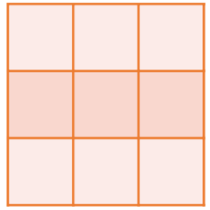
Auto-encoder



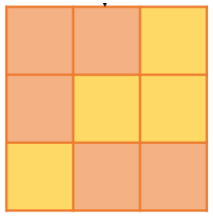
Framework

A cat in the snow → Text-to-image Generator → 

A cat in the snow → 1 Text Encoder → 



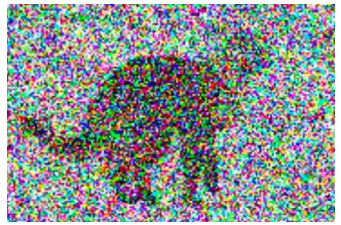
Generation Model 2

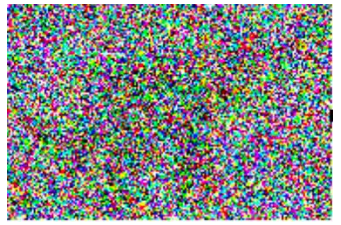


Decoder 3

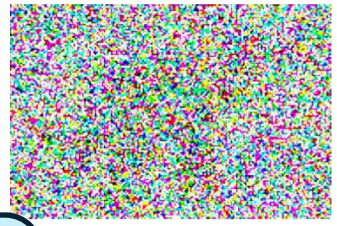
A cat in the snow

...

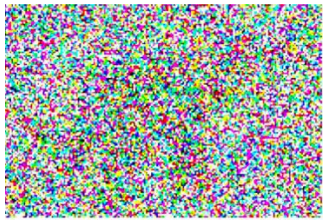




+



+



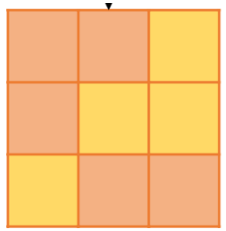
+

Encoder

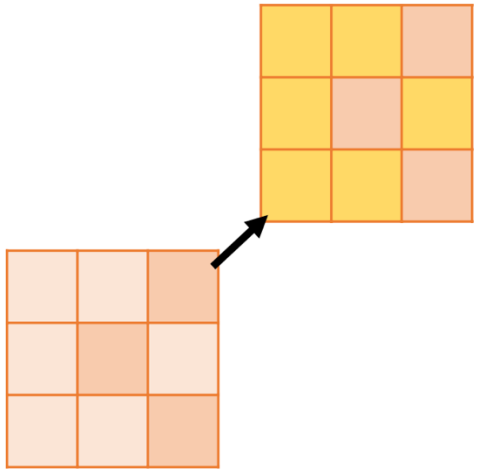
Step 1

Step 2

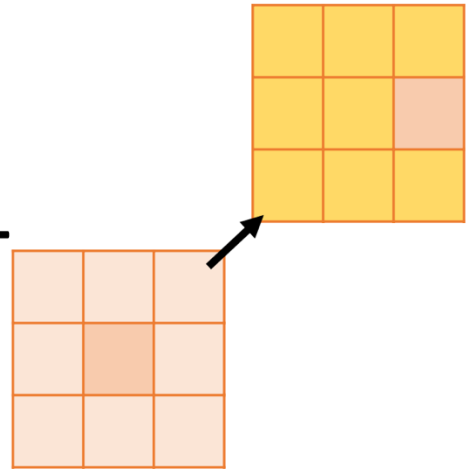
Step 1000



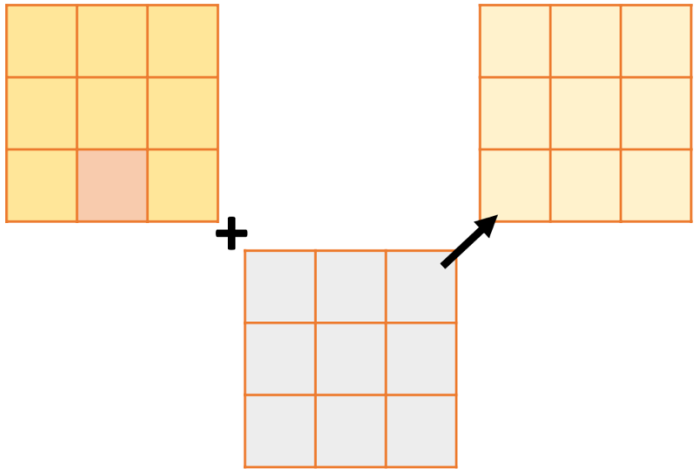
+



+

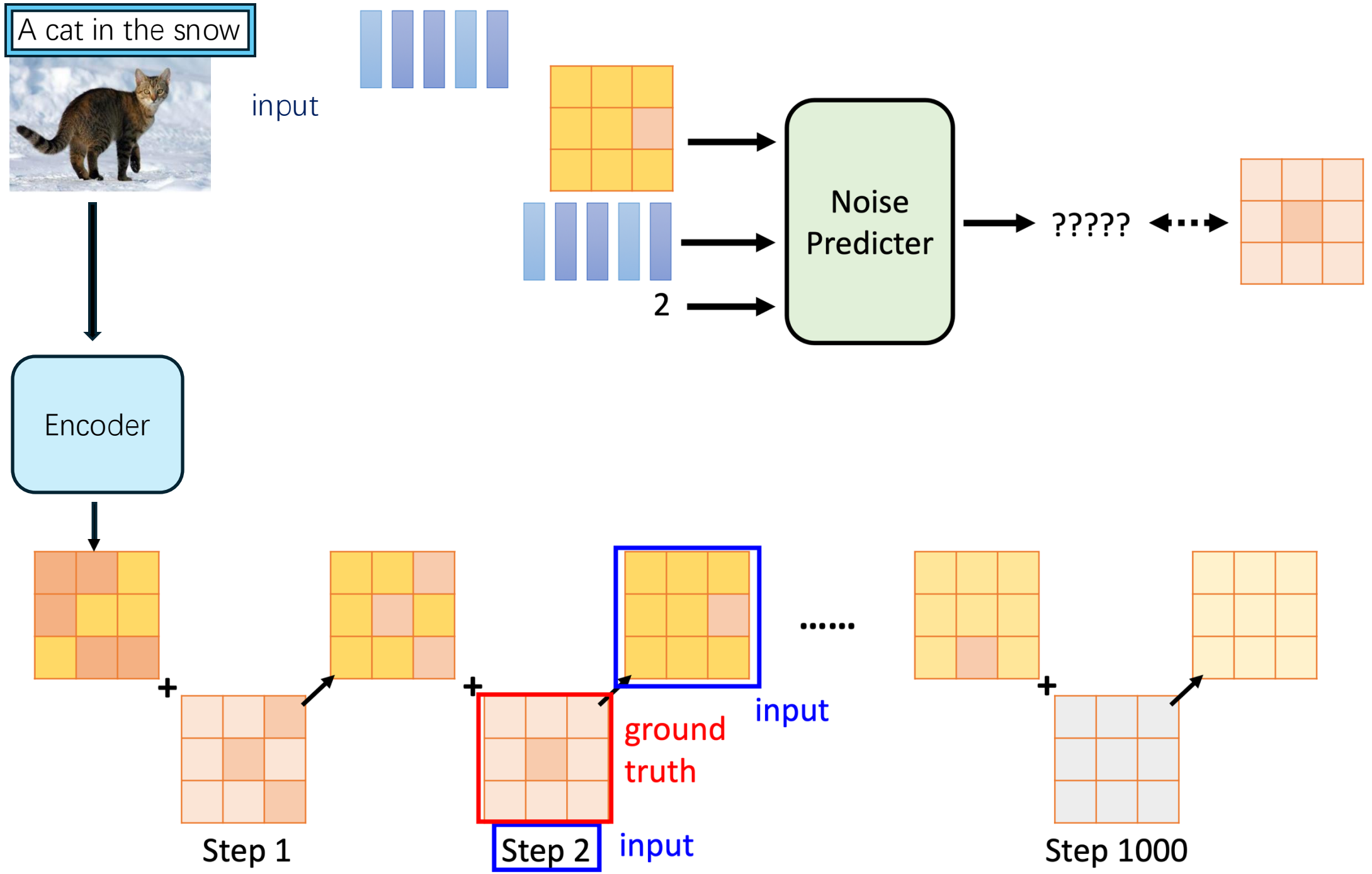


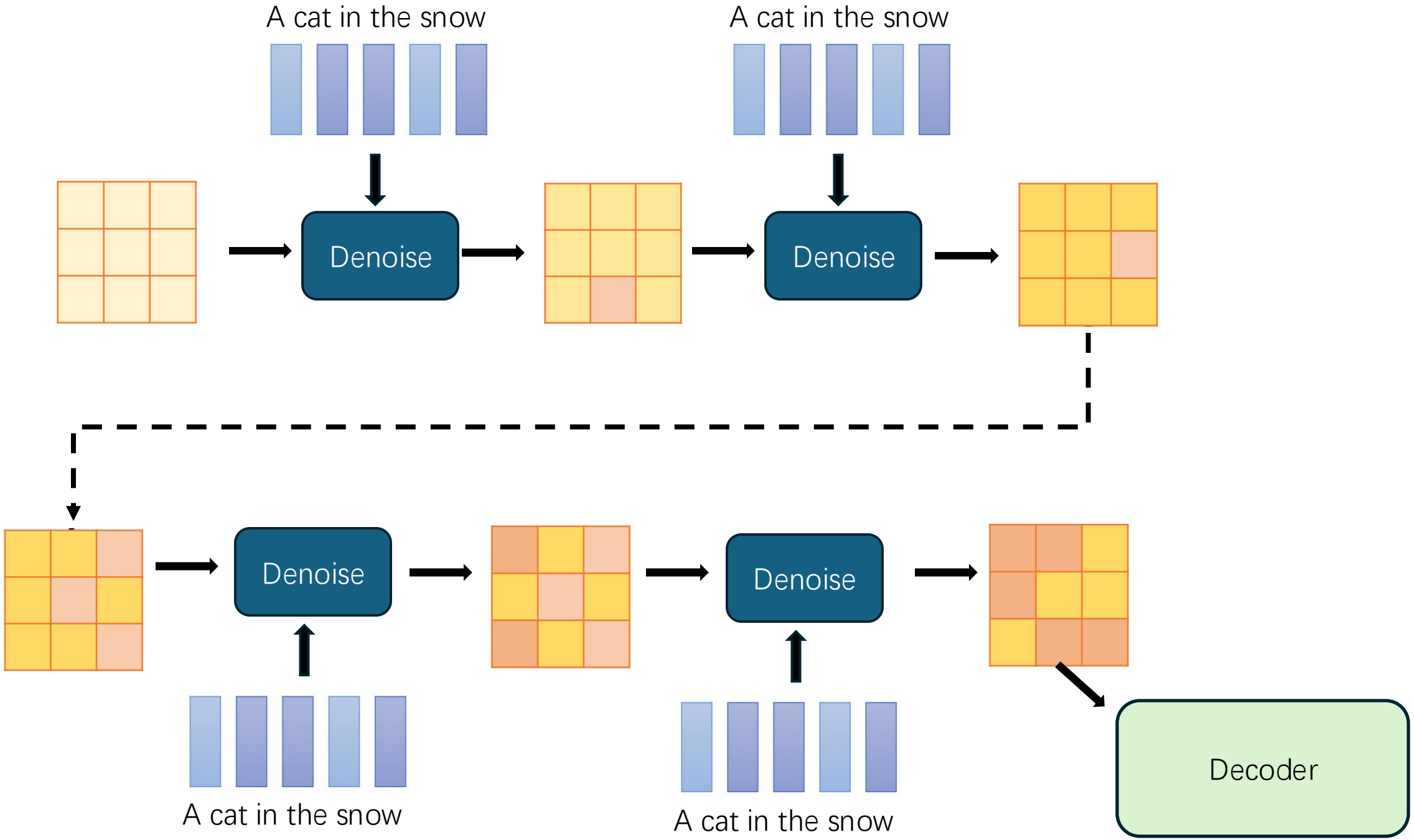
.....



+

noise

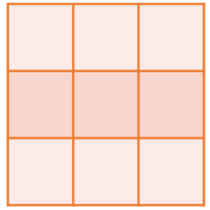
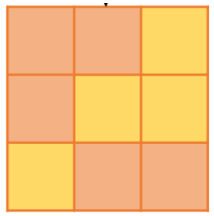

A cat in the snow
input
Noise Predicter
?????
Encoder
2
Step 1
Step 2
input
ground truth
input
.....
Step 1000


A cat in the snow
A cat in the snow
Denoise
Denoise
Denoise
Denoise
A cat in the snow
A cat in the snow
Decoder

Stable Diffusion

A cat in the snow → Text-to-image Generator → 

A cat in the snow → 1 Text Encoder → 

 → Generation Model 2 →  → Decoder 3 →