

JetFormer: An Autoregressive Generative Model of Raw Images and Text

Michael Tschannen, Andre Susano Pinto, Alexander Kolesnikov

Google DeepMind

ICLR 2025


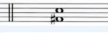
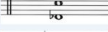
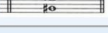

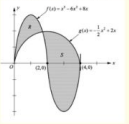
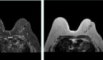
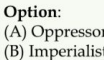


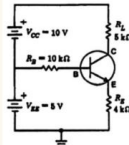
Presented by:

Nathan de Lara and Jasper Gerigk

March 28, 2025

Vision Language Models (VLM)

Learn Joint Distribution over Images and Text

<p>Art & Design</p> <p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <p>(A) Major third </p> <p>(B) Diminished fifth </p> <p>(C) Minor seventh </p> <p>(D) Diminished sixth </p> <p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Business</p> <p>Question: ...The graph shown is compiled from data collected by Gallup . Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <p>(A) 0 (B) 0.2142 (C) 0.3571 (D) 0.5</p> <p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Science</p> <p>Question:  The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <p>(A) $\int_0^{1.5} [f(x) - g(x)] dx$ (B) $\int_0^{1.5} [g(x) - f(x)] dx$ (C) $\int_0^2 [f(x) - g(x)] dx$ (D) $\int_0^2 [g(x) - x(x)] dx$</p> <p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>
<p>Health & Medicine</p> <p>Question: You are shown subtraction , T2 weighted  and T1 weighted axial  from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <p>(A) Susceptibility artifact (B) Hematoma (C) Fat necrosis (D) Silicone granuloma</p> <p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard</p>	<p>Humanities & Social Science</p> <p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? </p> <p>Option:</p> <p>(A) Oppressor (B) Imperialist (C) Savior (D) Isolationist</p> <p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Tech & Engineering</p> <p>Question: Find the VCE for the circuit shown in  Answer: 3.75</p> <p>Explanation: ...$I_E = [(V_{EE}) / (R_E)] = [(5 \text{ V}) / (4 \text{ k-ohm})] = 1.25 \text{ mA}$; $V_{CE} = V_{CC} - I_{ERL} = 10 \text{ V} - (1.25 \text{ mA}) 5 \text{ k-ohm}$; $V_{CE} = 10 \text{ V} - 6.25 \text{ V} = 3.75 \text{ V}$</p> <p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>

Source: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI ([link](#))

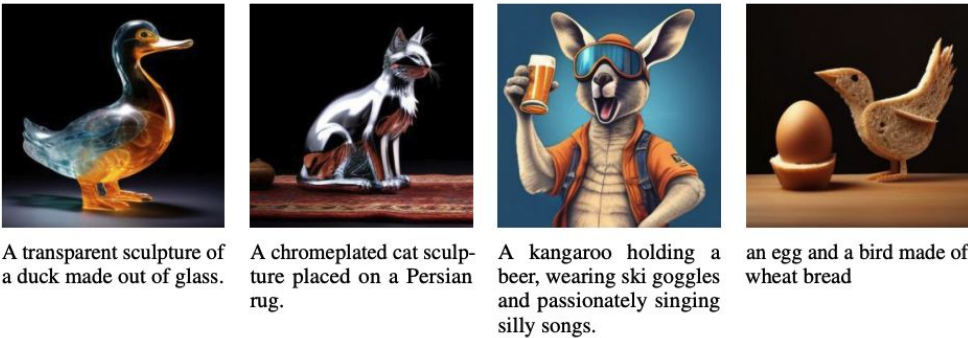


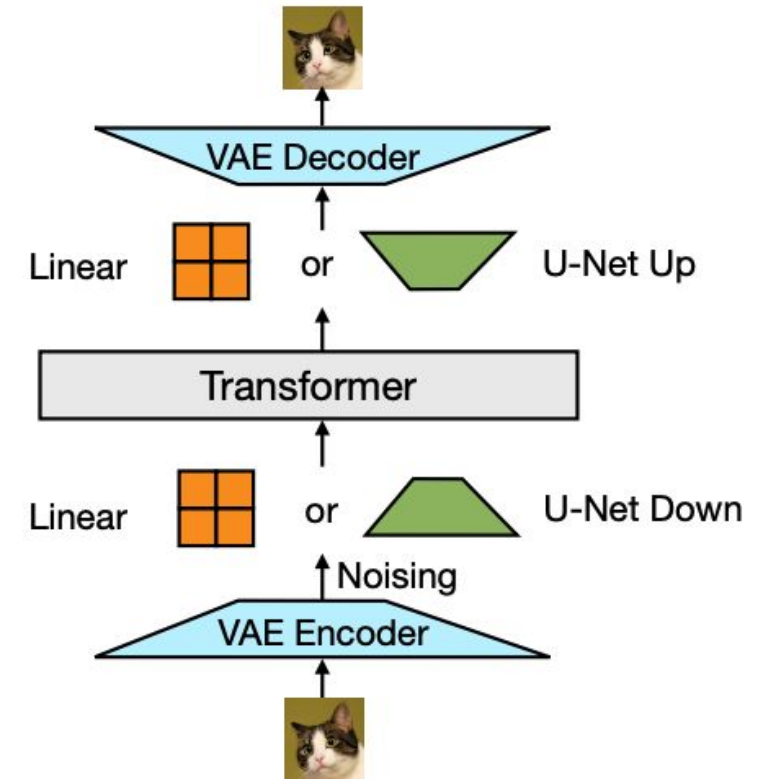
Figure 2: Generated images from a 7B Transfusion trained on 2T multi-modal tokens.

Source: Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model ([link](#))

Training an Image Generating VLM

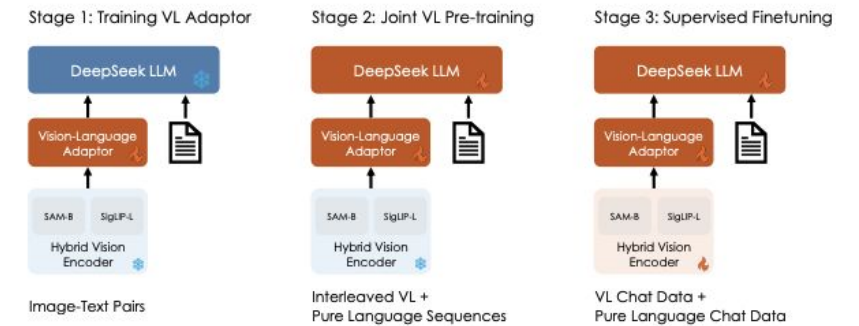
Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model

Chunting Zhou^{μ*} Lili Yu^{μ*} Arun Babu^{δ†} Kushal Tirumala^μ
Michihiro Yasunaga^μ Leonid Shamis^μ Jacob Kahn^μ Xuezhe Ma^σ
Luke Zettlemoyer^μ Omer Levy[†]

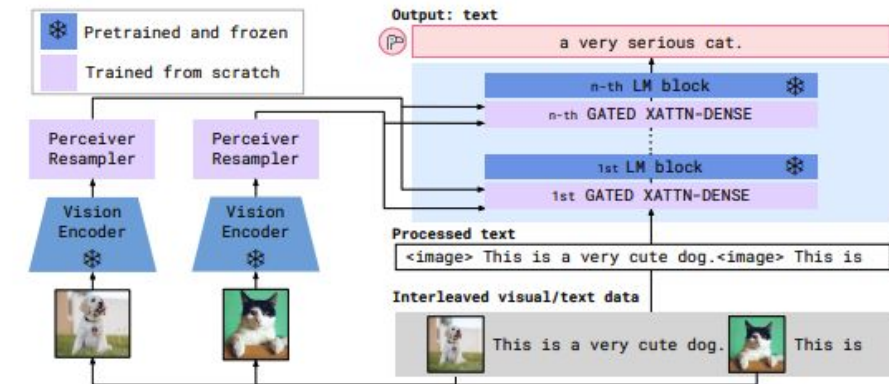


Training an Image Generating VLM

Popular existing methods freeze and train components separately during initial learning



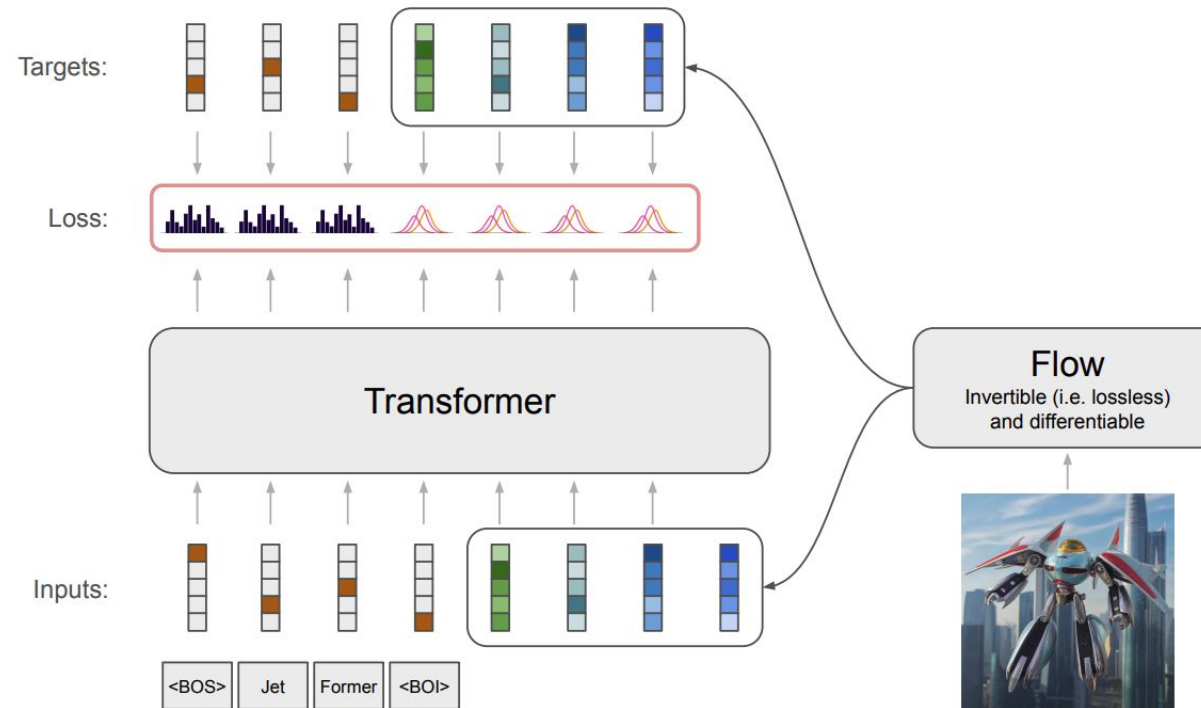
Source: DeepSeek-VL: Towards Real-World Vision-Language Understanding ([link](#))



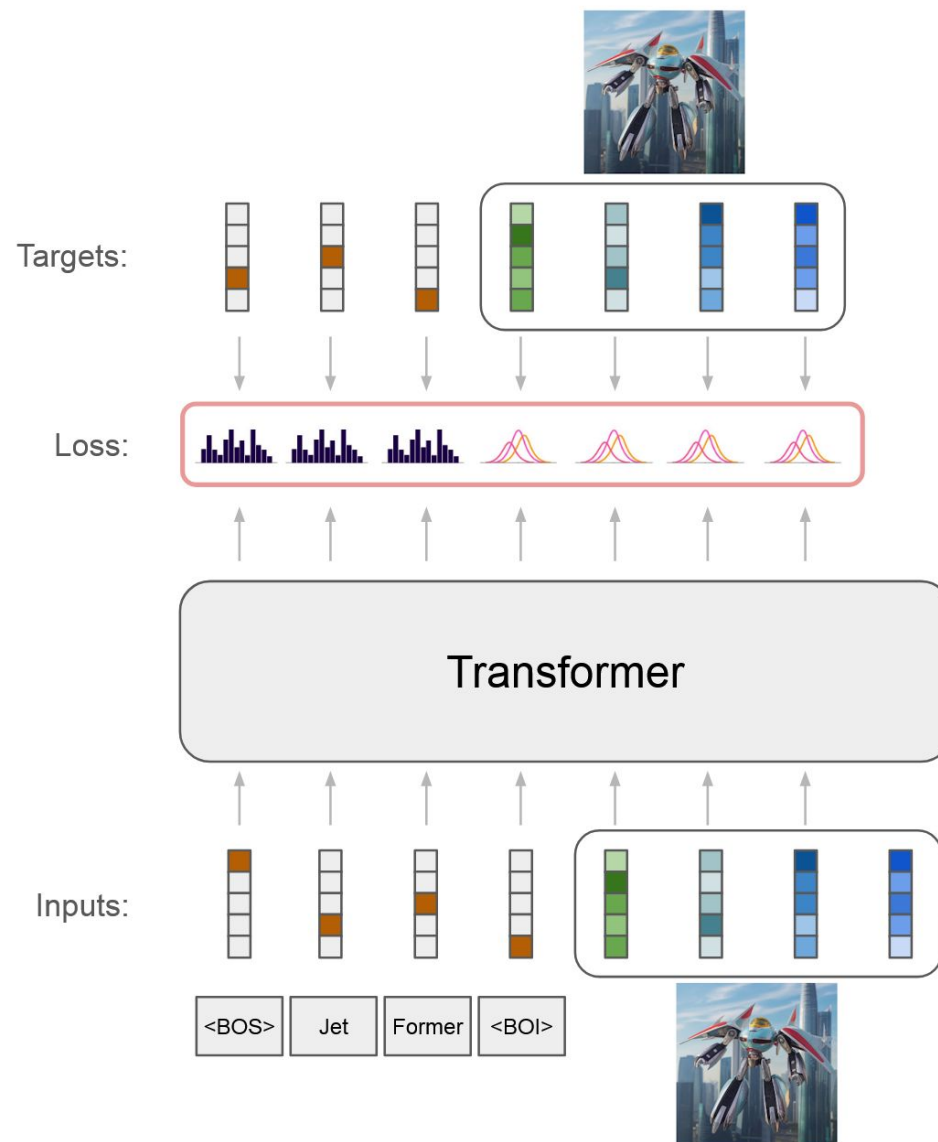
Source: Flamingo: A visual Language Model for Few-Shot Learning ([link](#))

Motivation

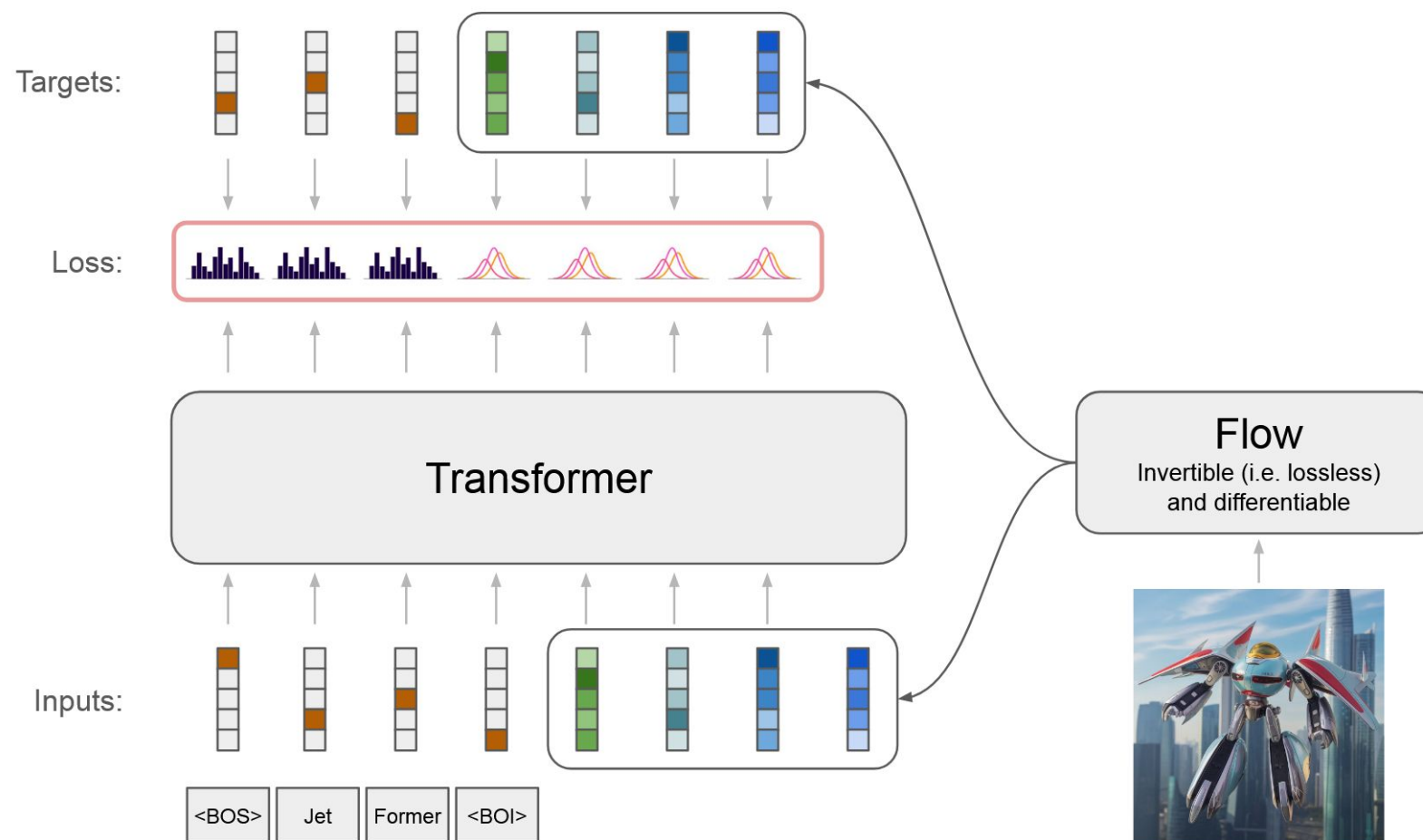
Replace freezing and individual component steps with one completely end-to-end trained model



JetFormer: An Autoregressive Generative Model of Raw Images and Text

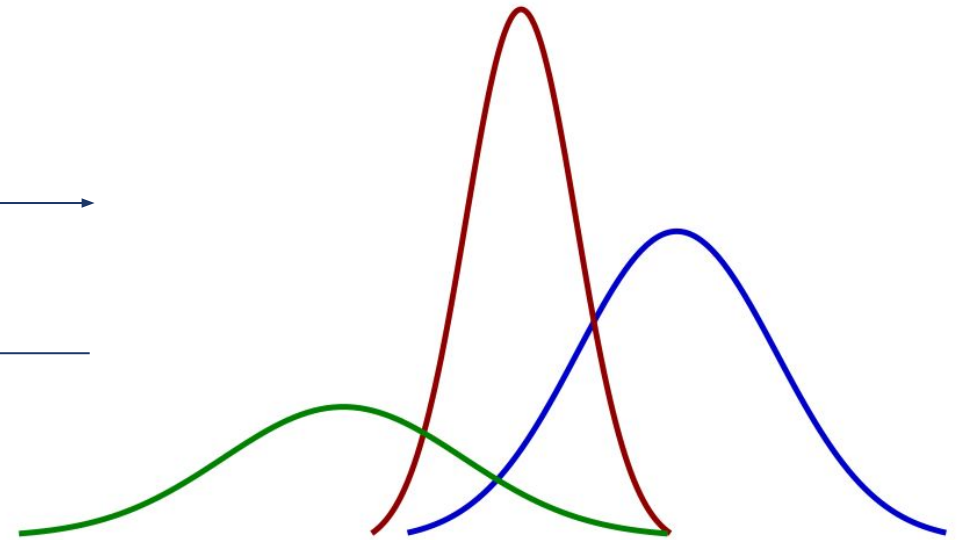
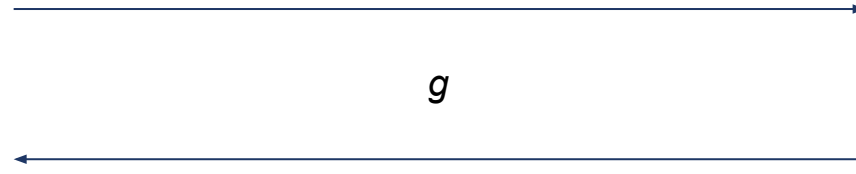


JetFormer: An Autoregressive Generative Model of Raw Images and Text



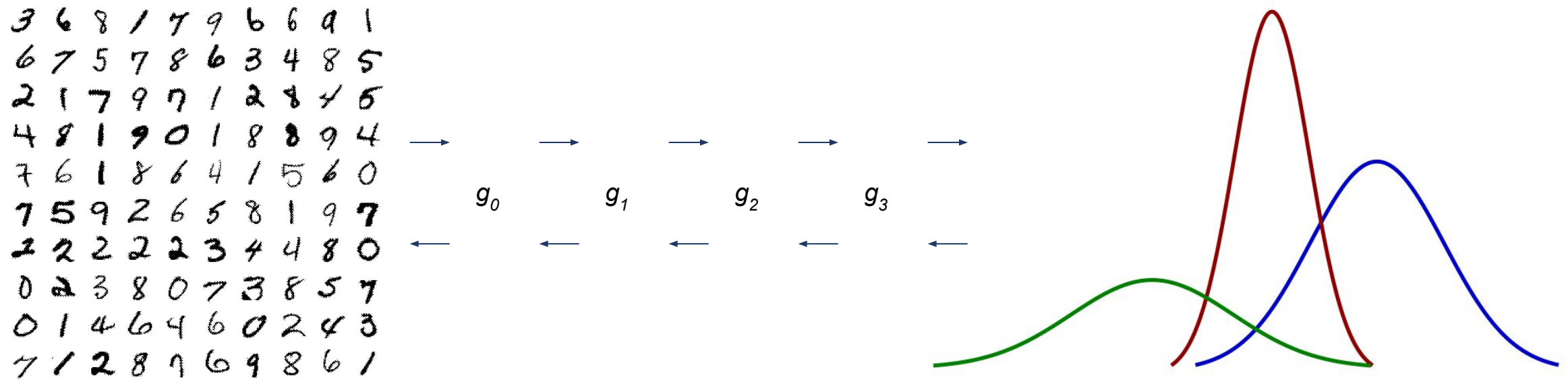
Flow Models

3 6 8 1 7 9 6 6 9 1
6 7 5 7 8 6 3 4 8 5
2 1 7 9 7 1 2 8 4 5
4 8 1 9 0 1 8 8 9 4
7 6 1 8 6 4 1 5 6 0
7 5 9 2 6 5 8 1 9 7
2 2 2 2 2 3 4 4 8 0
0 2 3 8 0 7 3 8 5 7
0 1 4 6 4 6 0 2 4 3
7 1 2 8 9 6 9 8 6 1



LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

Flow Models

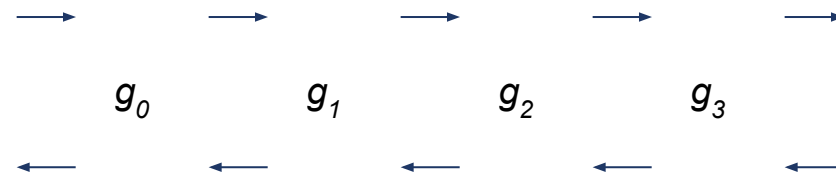


LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

Flow Models

3 6 8 1 7 9 6 6 9 1
 6 7 5 7 8 6 3 4 8 5
 2 1 7 9 7 1 2 8 4 5
 4 8 1 9 0 1 8 8 9 4
 7 6 1 8 6 4 1 5 6 0
 7 5 9 2 6 5 8 1 9 7
 2 2 2 2 2 3 4 4 8 0
 0 2 3 8 0 7 3 8 5 7
 0 1 4 6 4 6 0 2 4 3
 7 1 2 8 9 6 9 8 6 1

LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

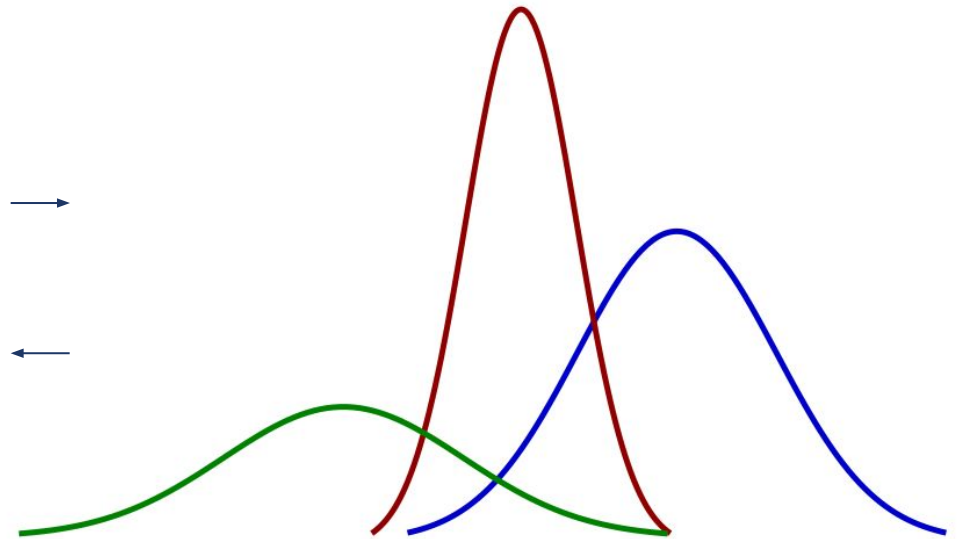


$$x = [x_1, x_2]$$

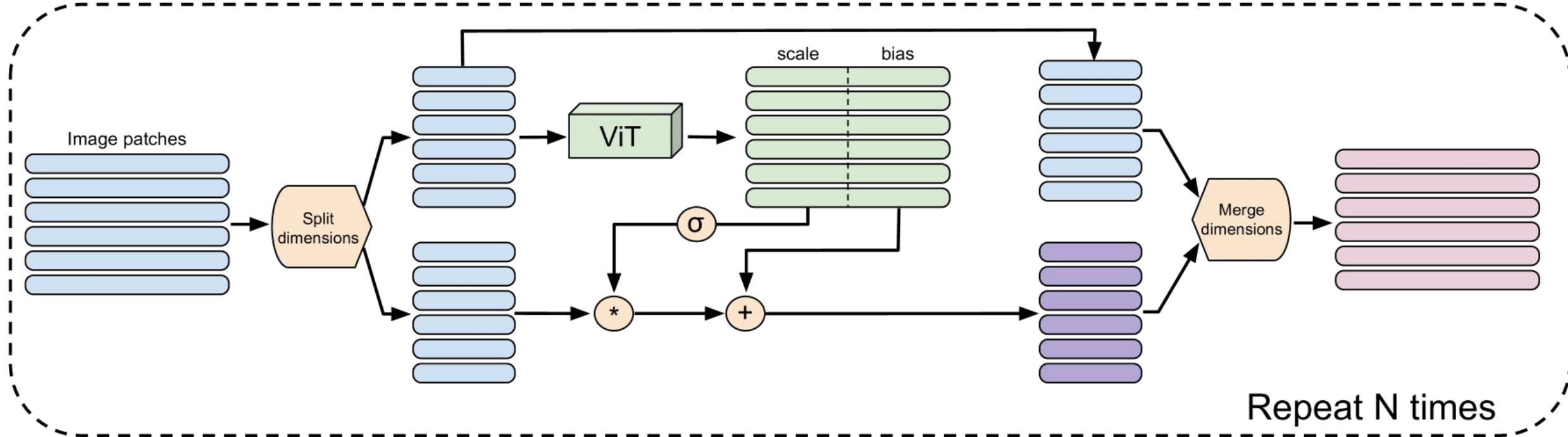
$$y_1 = x_1$$

$$y_2 = (x_2 + b(x_1)) \cdot \sigma(s(x_1)) \cdot m$$

Where b is the bias, s the scale,
and $m=2$ magnitude

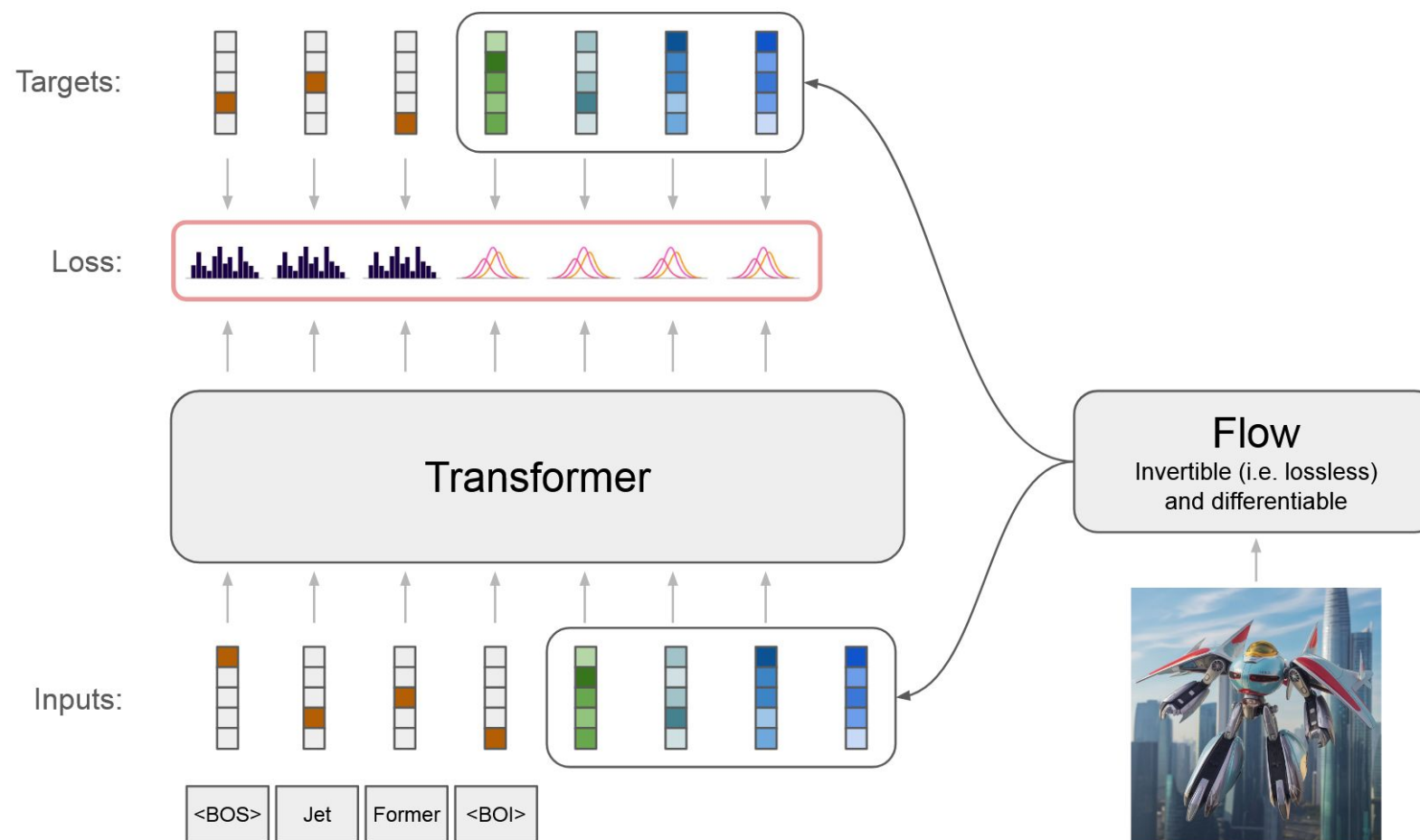


Jet: A Modern Transformer-Based Normalizing Flow



Kolesnikov, Alexander, André Susano Pinto, and Michael Tschannen. "Jet: A Modern Transformer-Based Normalizing Flow." *arXiv preprint arXiv:2412.15129* (2024).

JetFormer: An Autoregressive Generative Model of Raw Images and Text



Combining Auto-Regressive Transformer and Normalizing Flow

- **Use soft-tokens to represent the images**
 - Use the final hidden layer as the image representation
 - No loss of precision through discretization of tokens

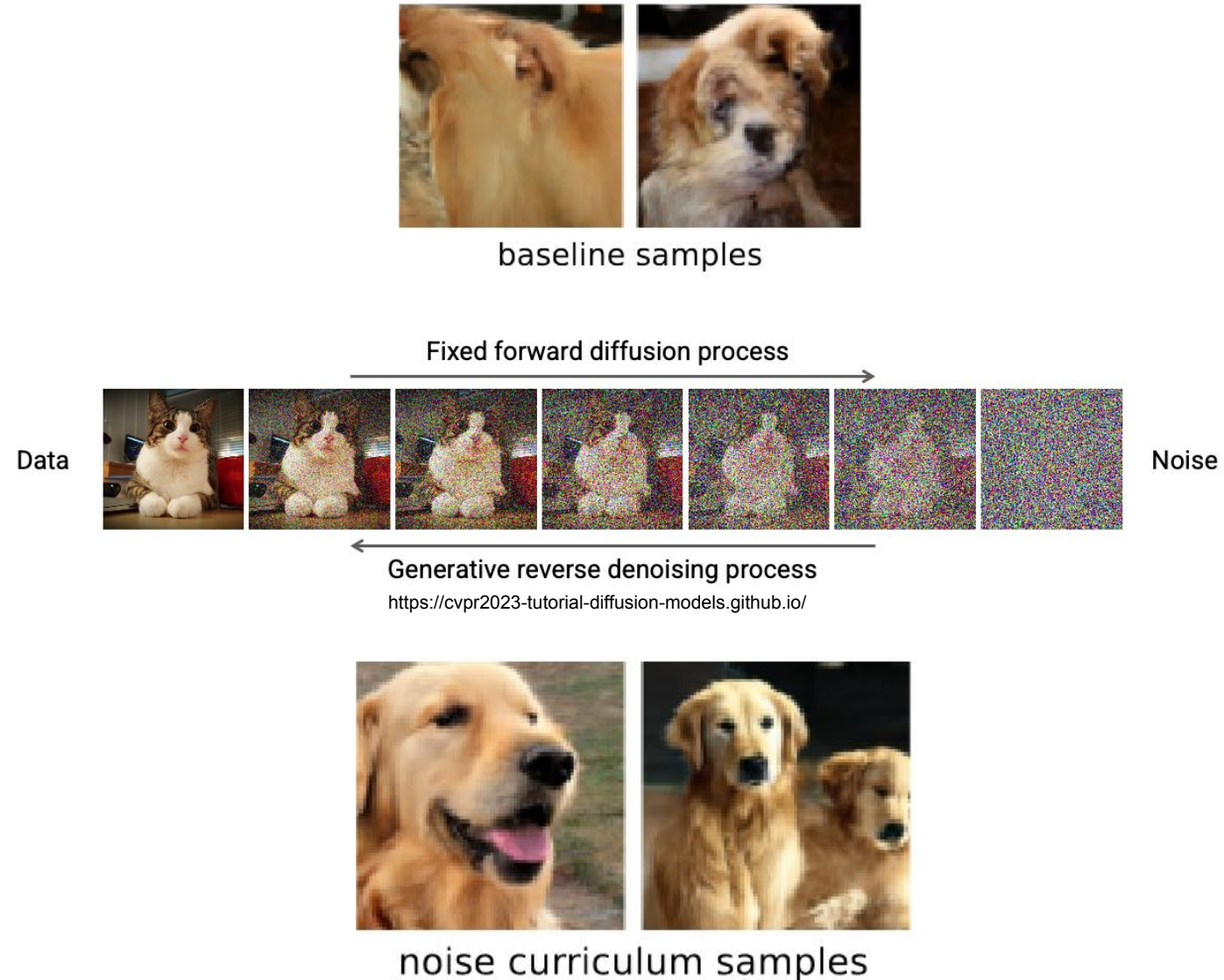
Combining Auto-Regressive Transformer and Normalizing Flow

- **Use soft-tokens to represent the images**
 - Use the final hidden layer as the image representation
 - No loss of precision through discretization of tokens
- **End-end learning unlocks full potential**
 - Auto-regressive predictions works better if key decisions are made first
 - End-end learning allows the order of channels to be influenced
 - Not possible when using a VAE (which must be pretrained)

Combining Auto-Regressive Transformer and Normalizing Flow

- **Use soft-tokens to represent the images**
 - Use the final hidden layer as the image representation
 - No loss of precision through discretization of tokens
- **End-end learning unlocks full potential**
 - Auto-regressive predictions works better if key decisions are made first
 - End-end learning allows the order of channels to be influenced
 - Not possible when using a VAE (which must be pretrained)
- **Normalizing Flow maintains number of dimensions**
 - This makes the task harder for the auto-regressive transformer
 - Model some of the channels using Gaussians and only some using the Flow
 - Same ideas as in probabilistic PCA

RGB Noise Curriculum During Training



Model Training

Normalizing Flows are Invertible  **Closed form available for log-likelihood!**

Change of variables formula: let X be a random variable with density $f_X(x)$ and g be a monotone function. Then density of the random variable $Y = g(X)$ is given by

$$f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X(g^{-1}(y)) .$$

Model Training

Normalizing Flows are Invertible  **Closed form available for log-likelihood!**

Train on (x, y) pairs and maximize log-likelihood of $p(y | x)$

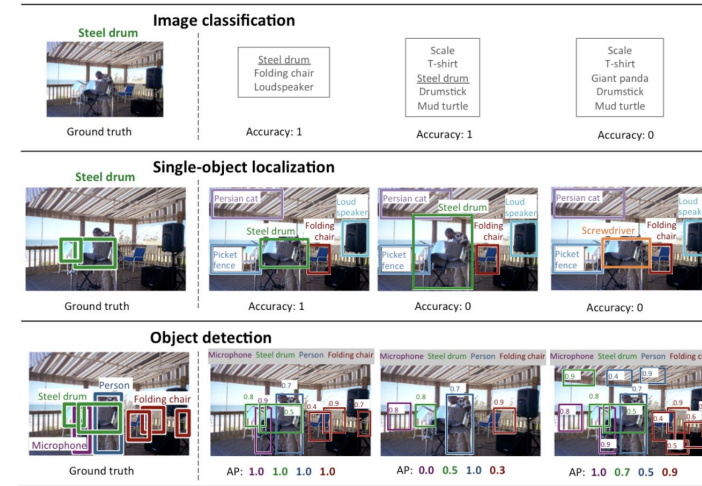
Maintain separate heads for text and image prediction

(x, y) pairs are (text, image) or (image, text)

Experiments

Training

- Text-to-Image
 - Training: ImageNet1K
 - Testing: Common Objects in Context (MS-COCO)
- Image-to-Text:
 - Training: WebLI dataset
 - Testing: ImageNet



Input: Generate the alt_text in EN
Output: A cellar filled with barrels of wine



Input: Generate the alt_text in EN
Output: a clock on a building that says 'sylvania' on it



Input: Generate the alt_text in EN
Output: Two helicopters are flying in the sky and one has a yellow stripe on the tail

Experimental Results

Text-to-Image Generation

	extra step	#param.	FID	FID (ft.)	NLL (ft.)
DALL-E (Ramesh et al., 2021)	VQ-VAE	12B	27.50		
CogView (Ding et al., 2021)	VQ-VAE	4B	27.10		
CogView2 (Ding et al., 2022)	VQ-VAE	6B	24.00	17.50	
ARGVLT (T&I) (Kim et al., 2023)	VQ-VAE	0.45B	16.93		
MAGVLT (T&I) (Kim et al., 2023)	VQ-VAE	0.45B	12.08		
Make-A-Scene (Gafni et al., 2022)	VQ-VAE	4B	11.84	7.55	
LDM-KL-8-G (Rombach et al., 2022)	VAE	1.45B	12.63		
GLIDE (Nichol et al., 2022)	Super-res.	6B	12.24		
DALL-E-2 (Ramesh et al., 2022)	Super-res.	5.2B	10.39		
JetFormer-L (T&I)	–	2.75B	20.86	13.70	3.86
JetFormer-L	–	2.75B	18.63	13.07	3.85

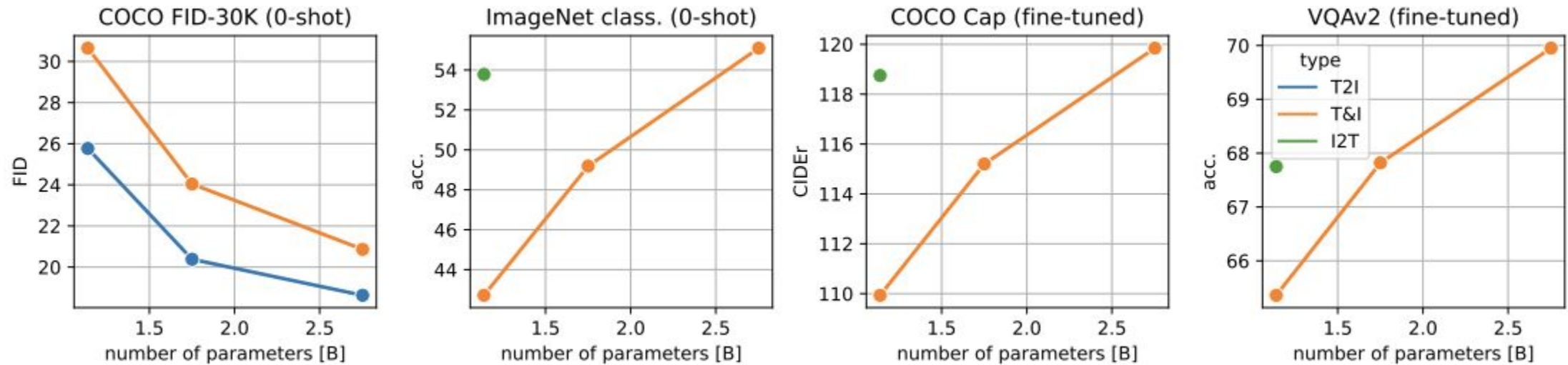
Experimental Results

Image-to-Text Understanding (VQA)

	extra step	COCO cap.	VQAv2
CapPa L/14 (Tschannen et al., 2023)*	–	118.7	68.6
CLIP L/14 (Radford et al., 2021)*	–	118.2	67.9
ARGVLT (T&I) (Kim et al., 2023)	VQ-VAE	94.7	–
MAGVLT Large (T&I) (Kim et al., 2023)	VQ-VAE	110.7	65.7
JetFormer-B (I2T)	–	118.7	67.2
JetFormer-L (T&I)	–	119.8	70.0

Experimental Results

Scaling Dynamics



Limitations & Future Work

- Final trained model often fails to out-perform baselines
- Limited training and evaluation
- End-to-End training also requires more compute since image encodings cannot be pre-calculated
- Separate heads for image and text means responses are modality restricted

Conclusions & Future Work

- End-to-End Text and Image models are trainable by stitching together existing architectures!
- Optimizing closed-form log-likelihood enables end-to-end objective
- Adding noise and gradually removing throughout training is crucial

Thank you!

Now to the colab
After: Questions?



UNIVERSITY OF
TORONTO

DEFY
GRAVITY