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

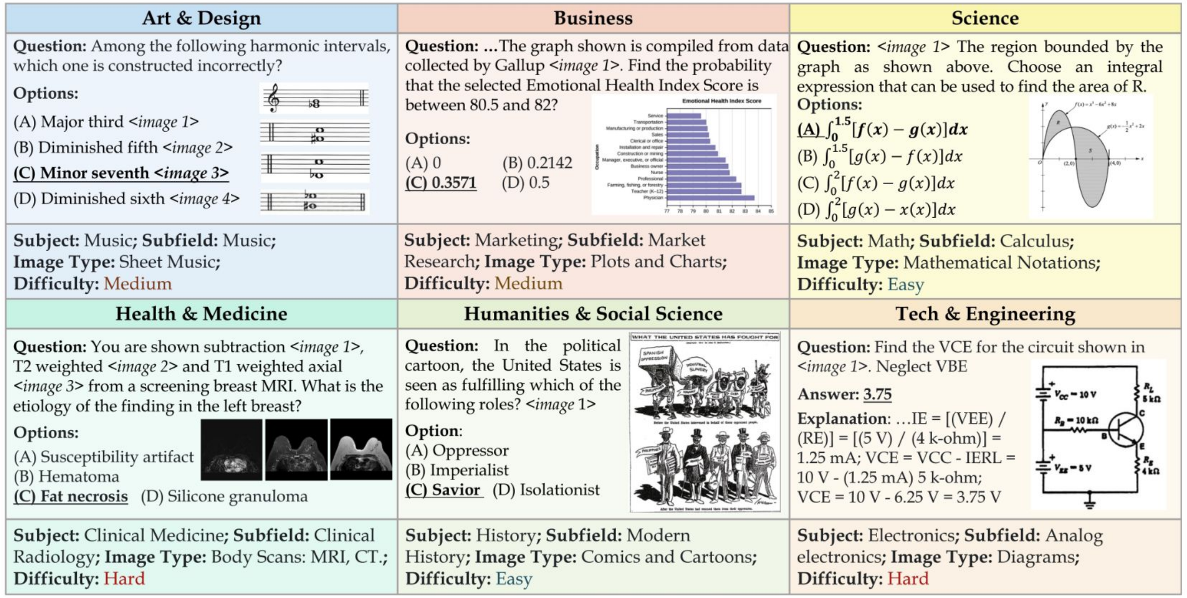
March 28, 2025

UNIVERSITY OF TORONTO

DEFY GRAVITY

Vision Language Models (VLM)

Learn Joint Distribution over Images and Text



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

A transparent sculpture of a duck made out of glass. | A chromeplated cat sculpture placed on a Persian rug. | A kangaroo holding a beer, wearing ski goggles and passionately singing silly songs. | an egg and a bird made of wheat bread

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

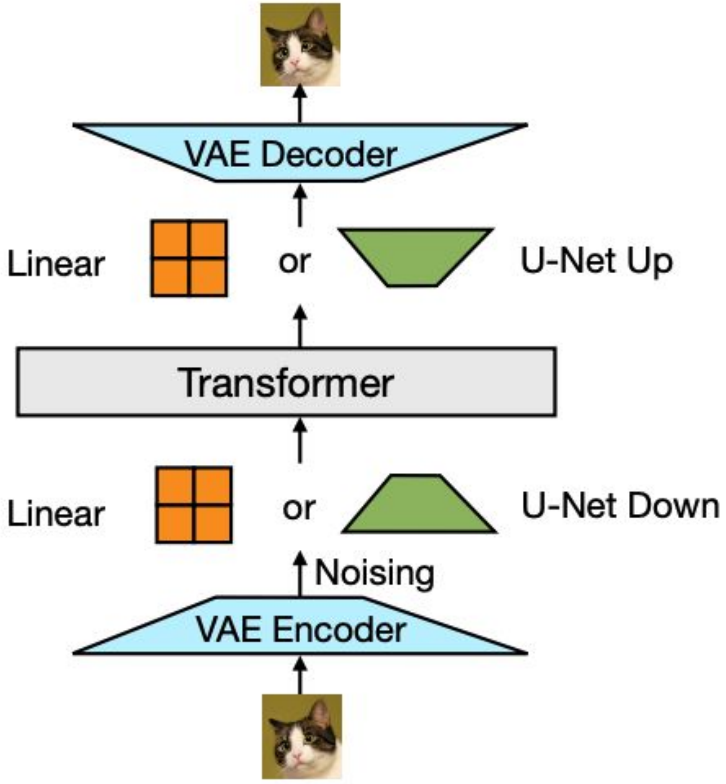
Source: Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model (link)

UNIVERSITY OF TORONTO

Training an Image Generating VLM

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

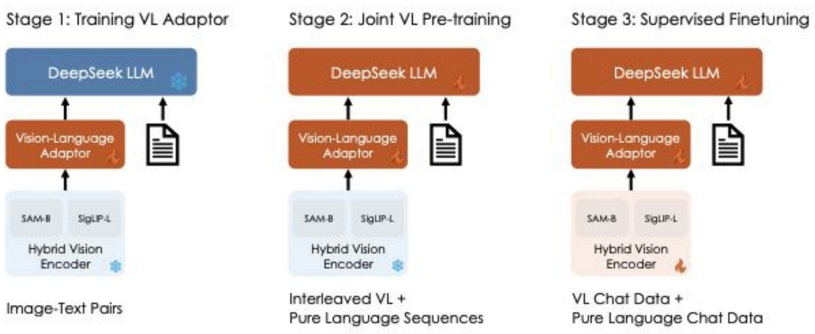
Chunting Zhou[μ*] Lili Yu[μ*] Arun Babu[δ†] Kushal Tirumala[μ]
Michihiro Yasunaga[μ] Leonid Shamis[μ] Jacob Kahn[μ] Xuezhe Ma[σ]
Luke Zettlemoyer[μ] Omer Levy[†]



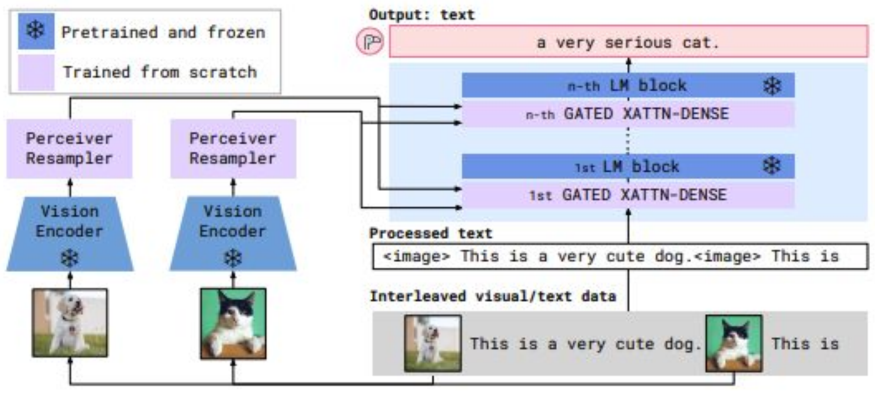
UNIVERSITY OF TORONTO

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)

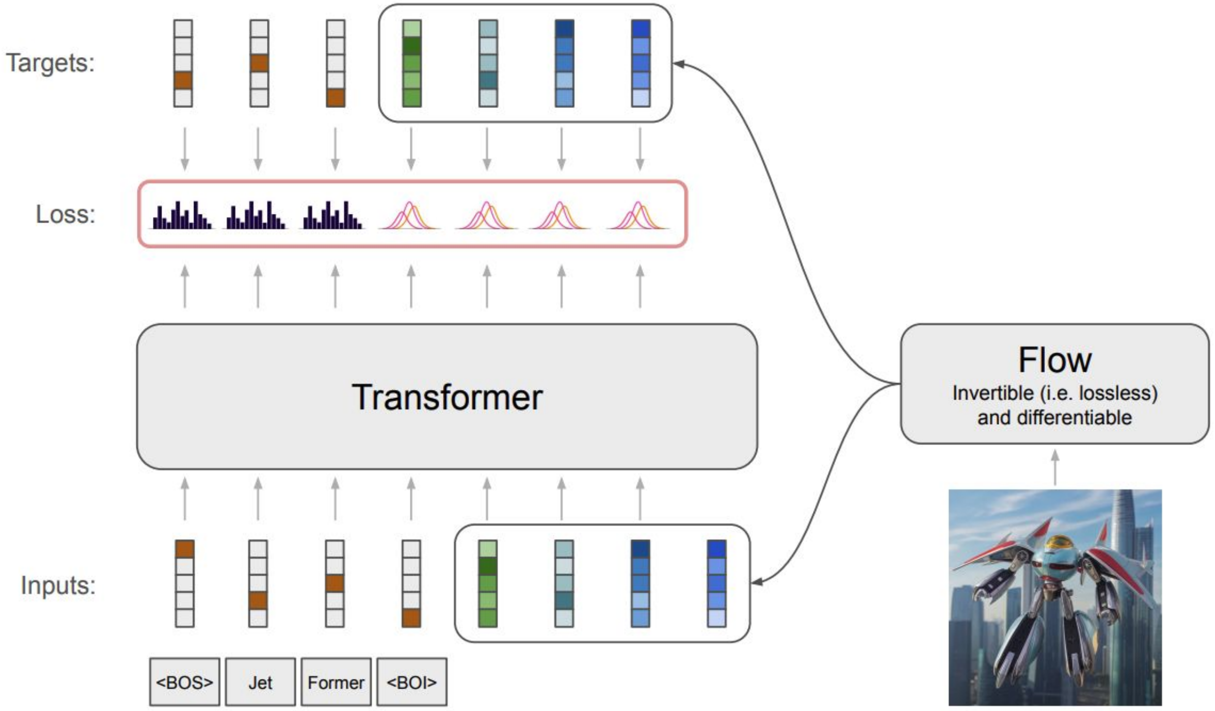


Source: Flamingo: A visual Language Model for Few-Shot Learning (link)

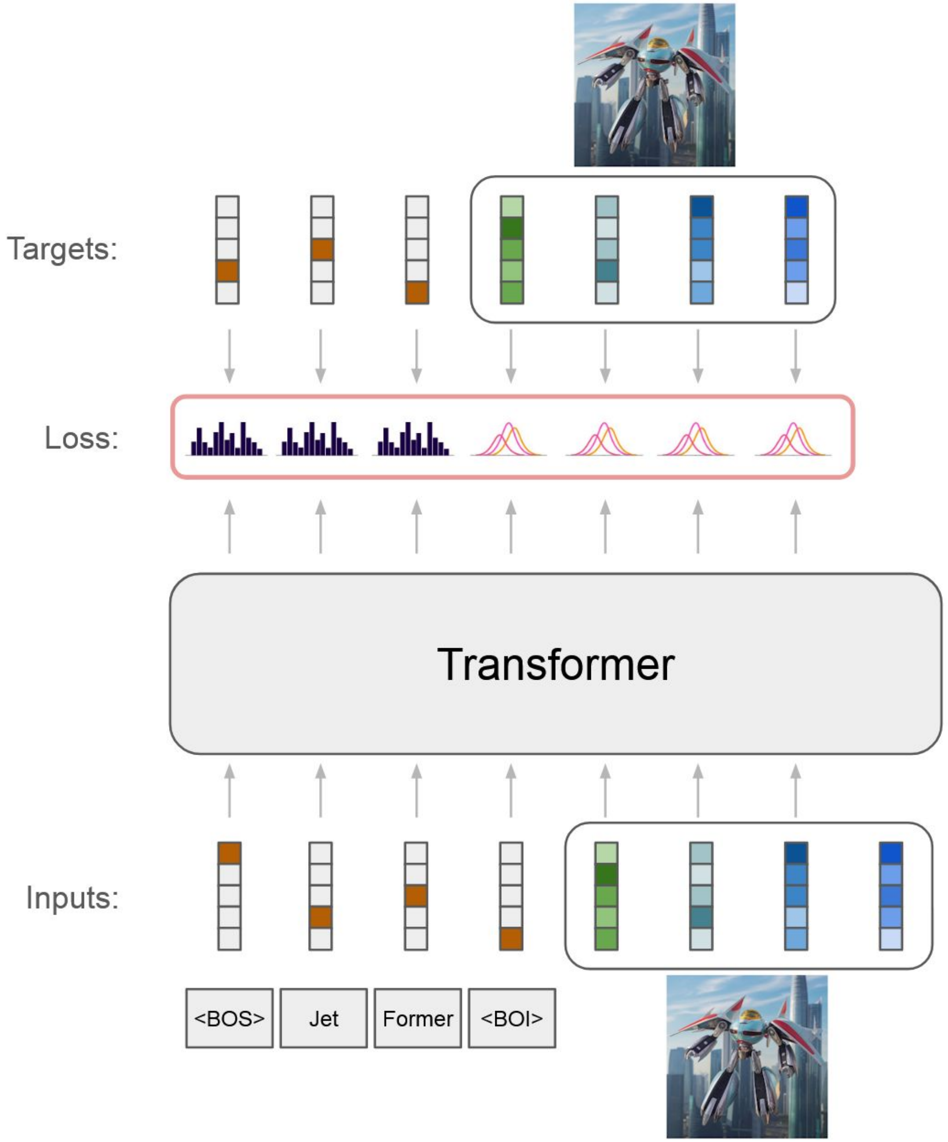
UNIVERSITY OF TORONTO

Motivation

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

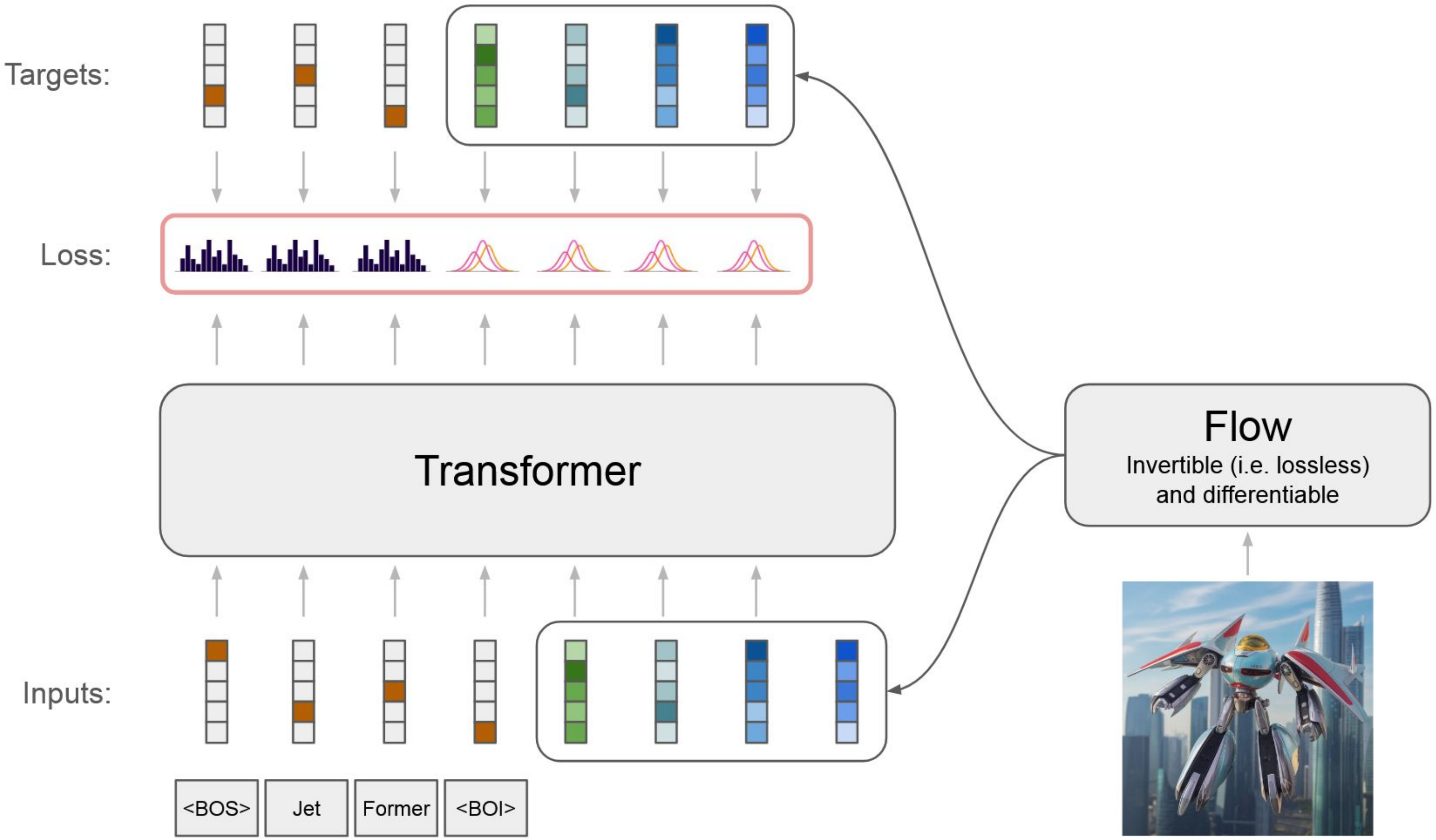


JetFormer: An Autoregressive Generative Model of Raw Images and Text



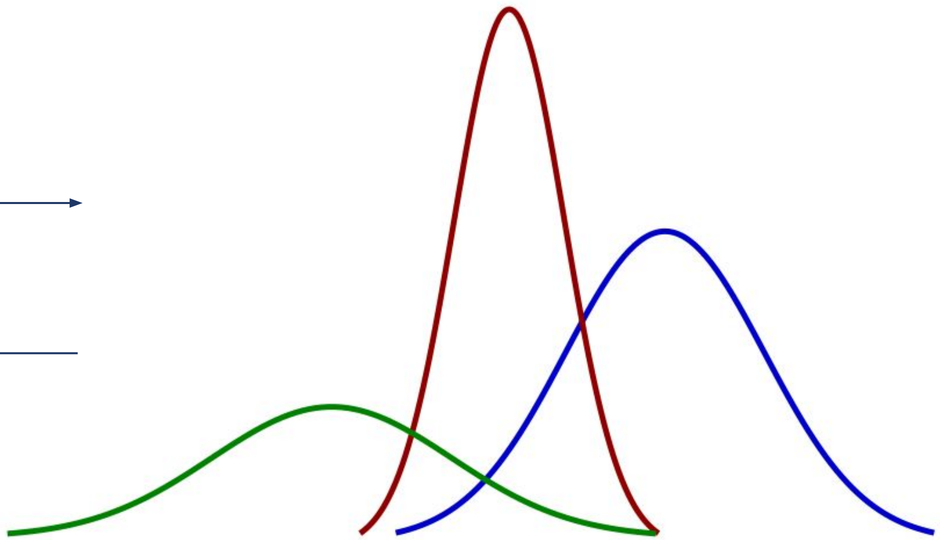
UNIVERSITY OF TORONTO

JetFormer: An Autoregressive Generative Model of Raw Images and Text



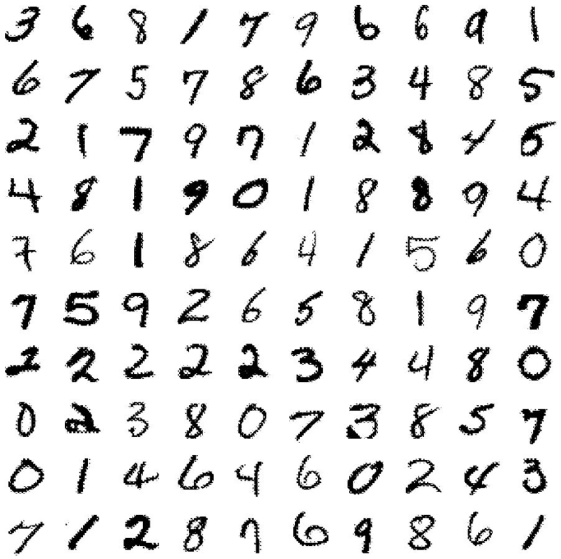
Flow Models

g

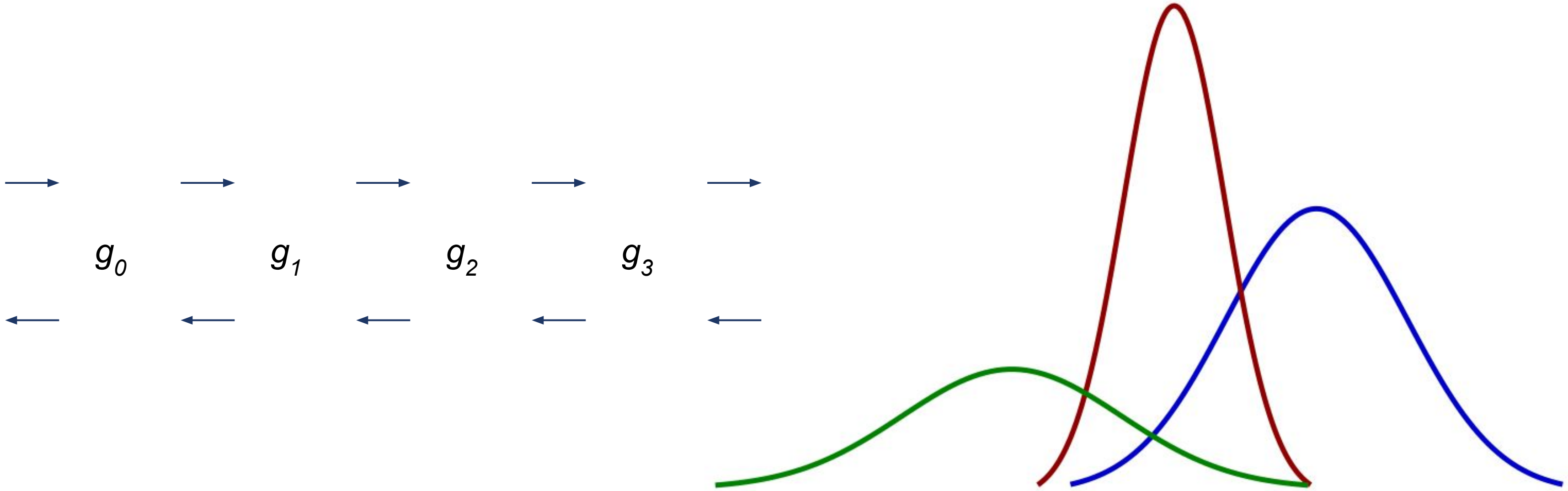


LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

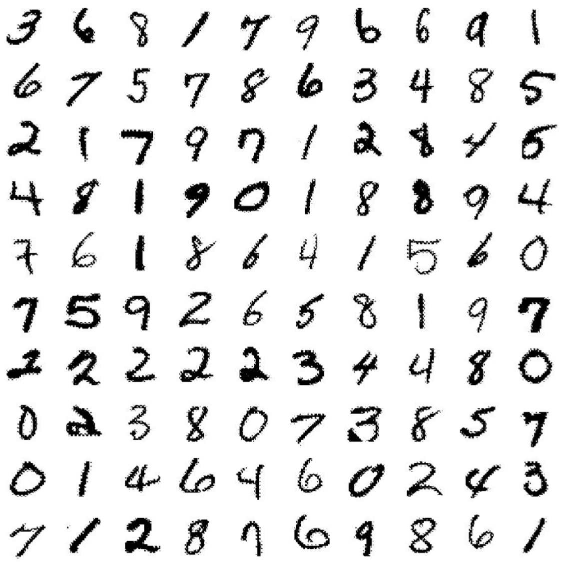
Flow Models



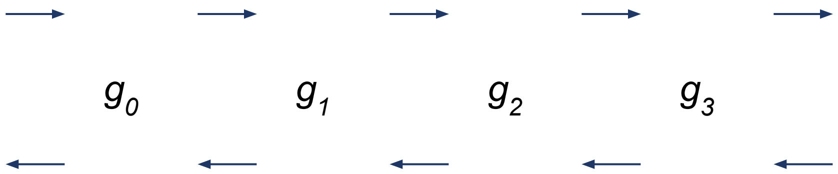
LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

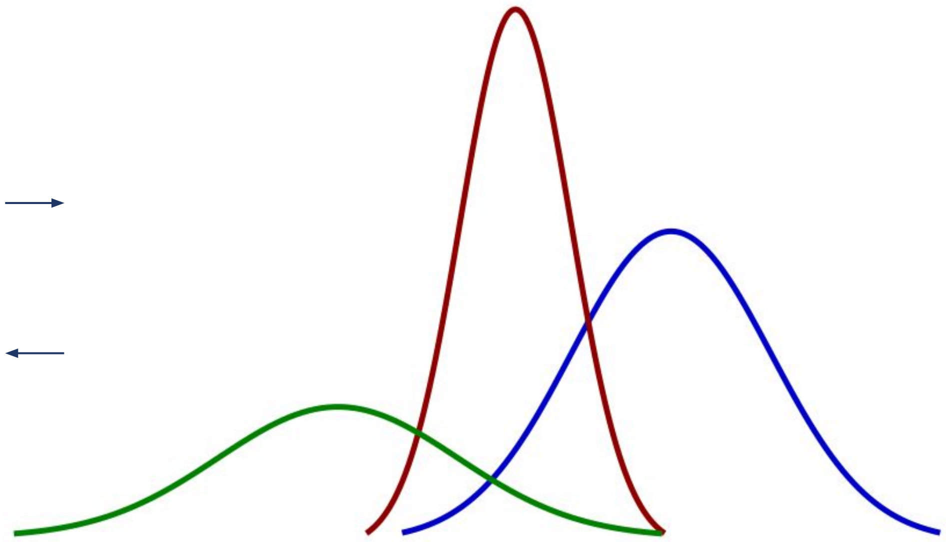


Flow Models



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

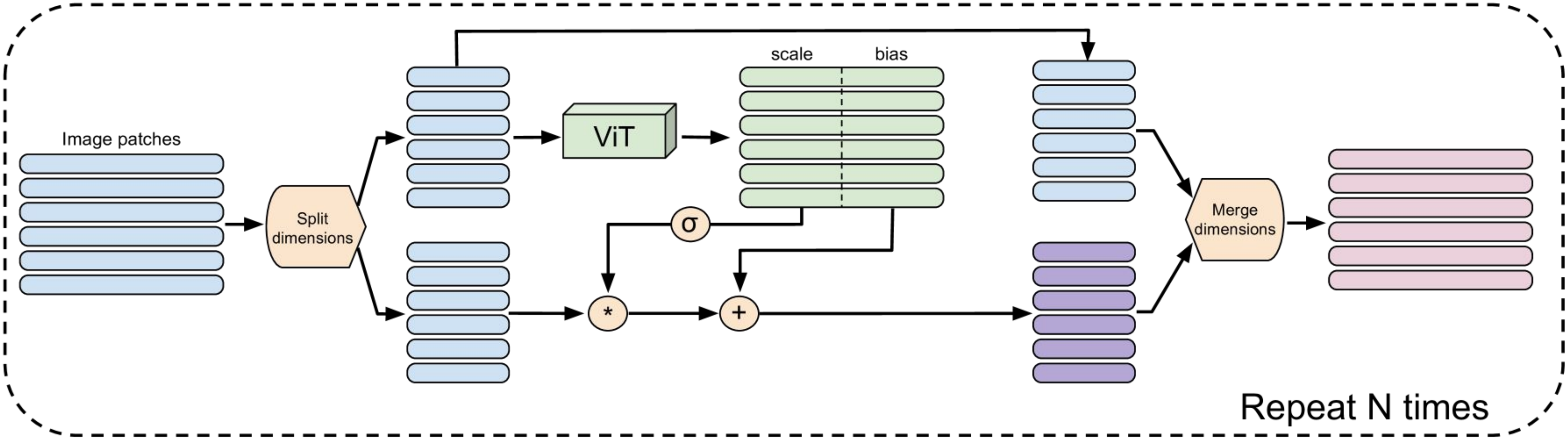




$$x = [x_1, x_2]$$
$$y_1 = x_1$$
$$y_2 = (x_2 + b(x_1)) \cdot \sigma(s(x_1)) \cdot m$$

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

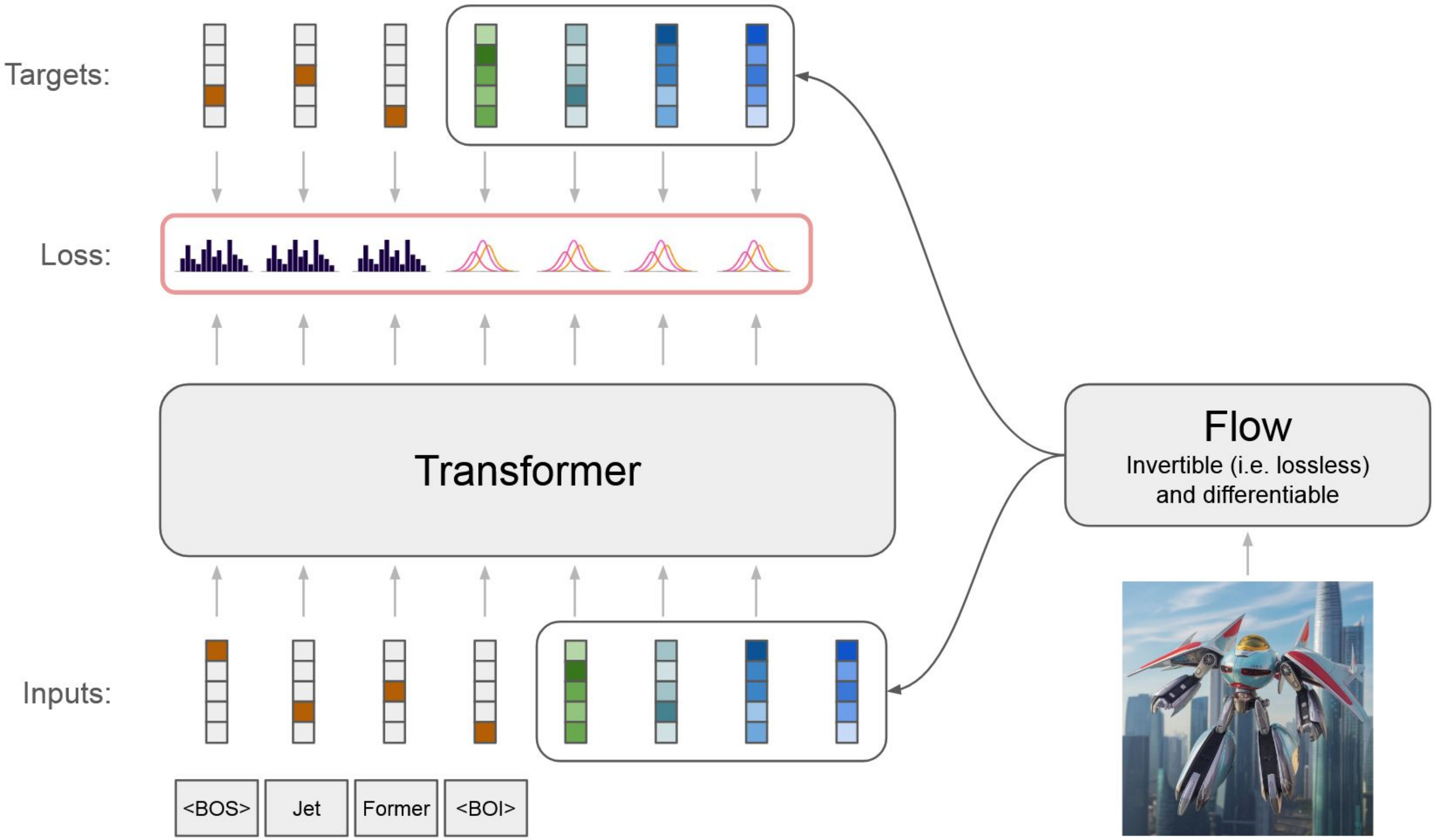
Jet: A Modern Transformer-Based Normalizing Flow



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

UNIVERSITY OF TORONTO

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)

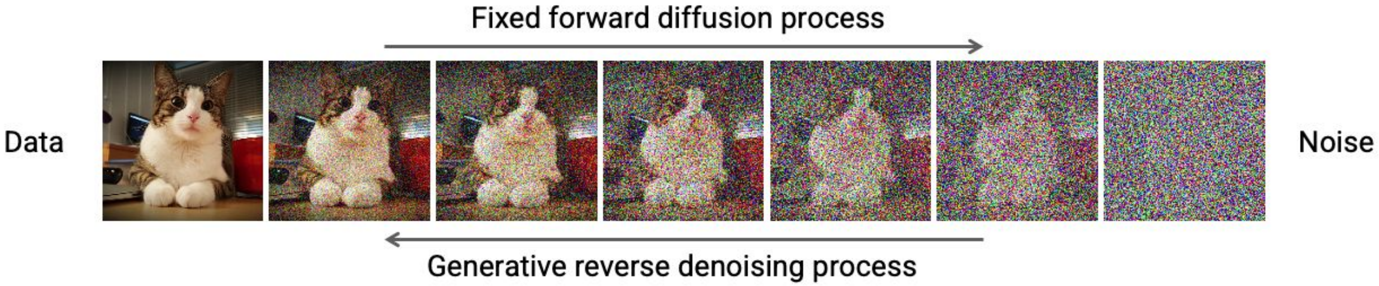
UNIVERSITY OF TORONTO

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

baseline samples



https://cvpr2023-tutorial-diffusion-models.github.io/

noise curriculum samples

UNIVERSITY OF TORONTO

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

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

Model Training

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

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

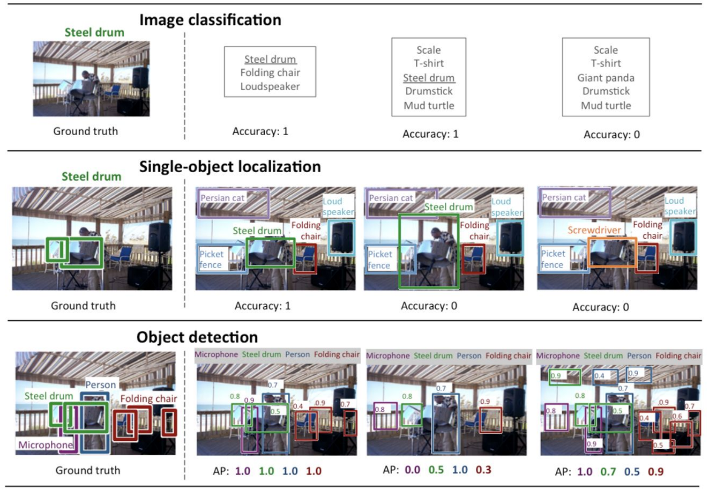
Maintain separate heads for text and image prediction

(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 'lyvania' 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

UNIVERSITY OF TORONTO

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

UNIVERSITY OF TORONTO

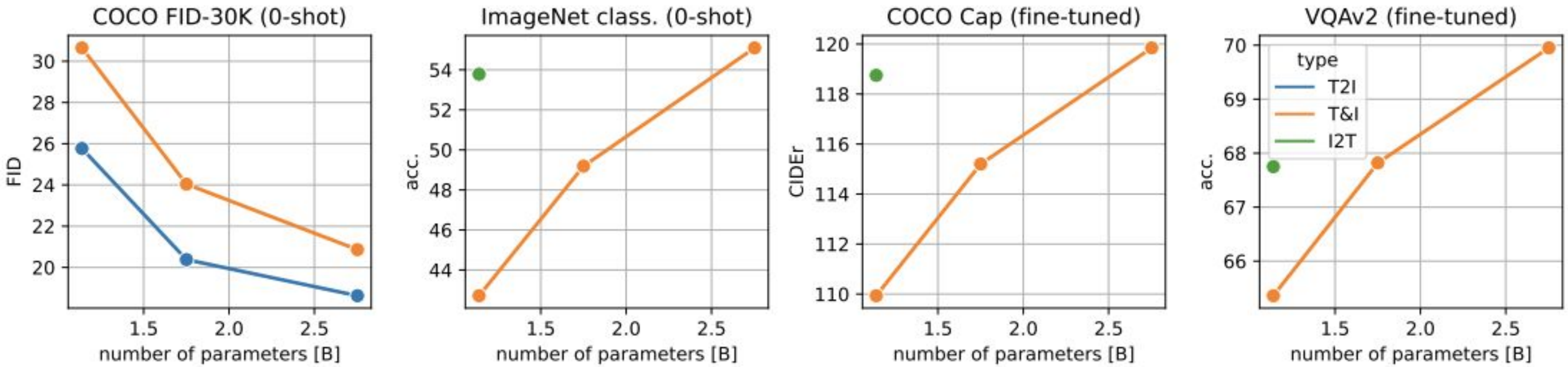
Experimental Results

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



UNIVERSITY OF TORONTO

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