JetFormer: An Autoregressive Generative Model of Raw Images and Text

Michael Tschannen* André Susano Pinto* Alexander Kolesnikov*,°
Google DeepMind
*equal contribution °work done while at Google DeepMind

Motivation

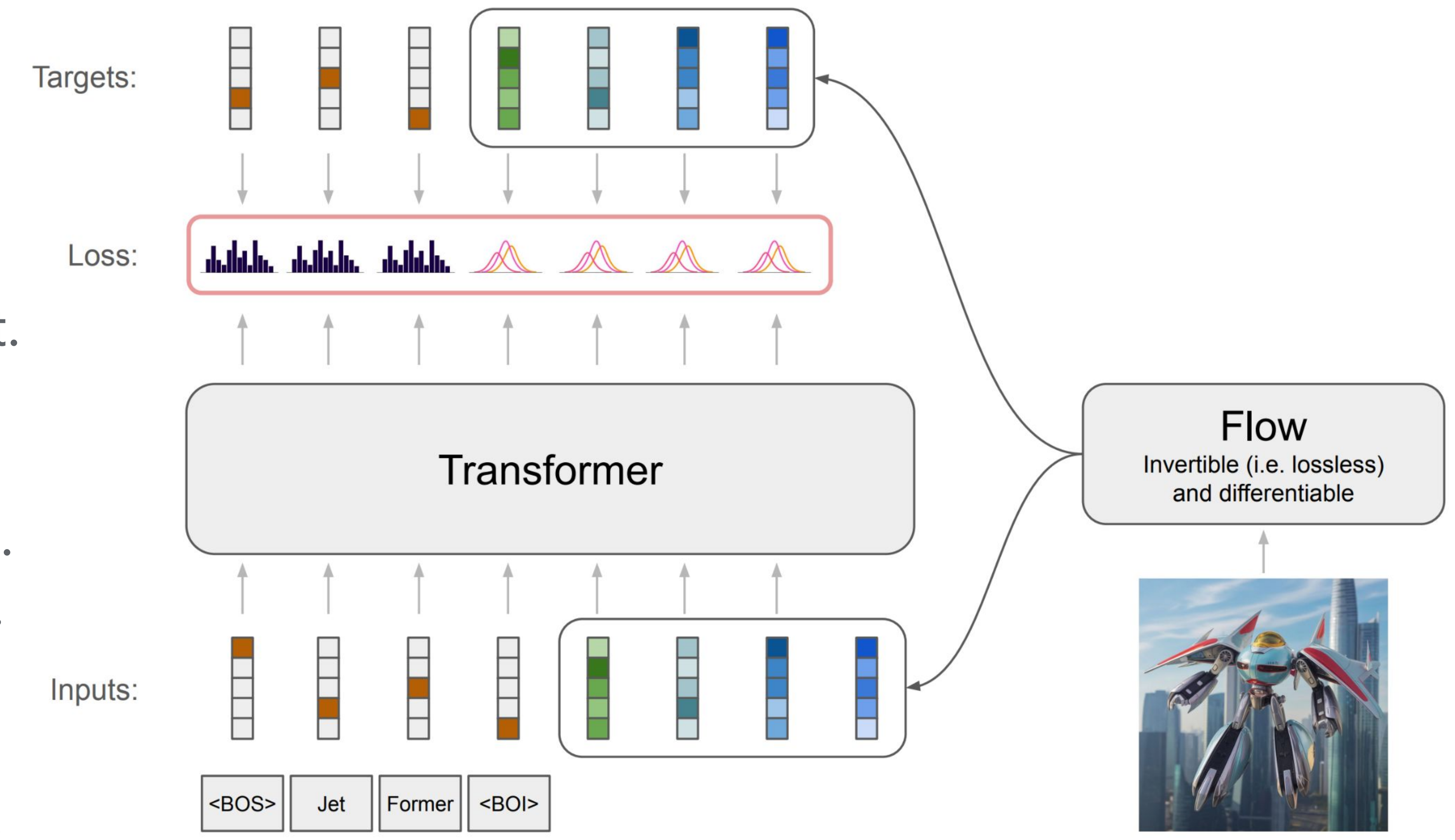
Removing modeling constraints and unifying architectures across domains has been a key driver of the recent progress in training large multimodal models. However, most of these models still rely on many separately trained components such as modality-specific encoders and decoders which can limit performance on certain tasks. For example, general-purpose (VQ-)VAEs for images can limit generalization to fine-grained dense prediction tasks due to their lossy latent representation.

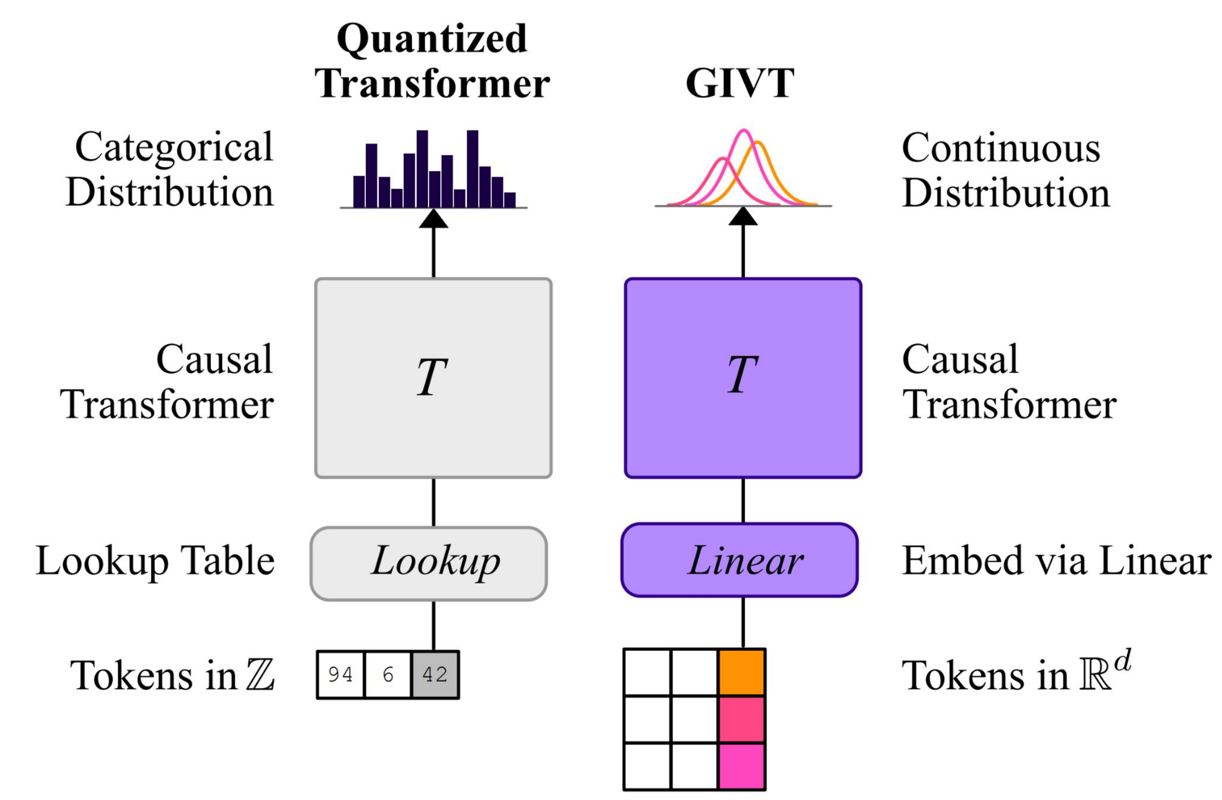
In this work, we further streamline joint generative modeling of images and text. **We propose an autoregressive decoder-only transformer—JetFormer—which is trained to directly maximize the likelihood of raw data, without relying on any separately pretrained components, and can understand and generate both text and images.** By design JetFormer relies on a lossless image representation and hence can overcome some of limitations of pretrained encoders/decoder.

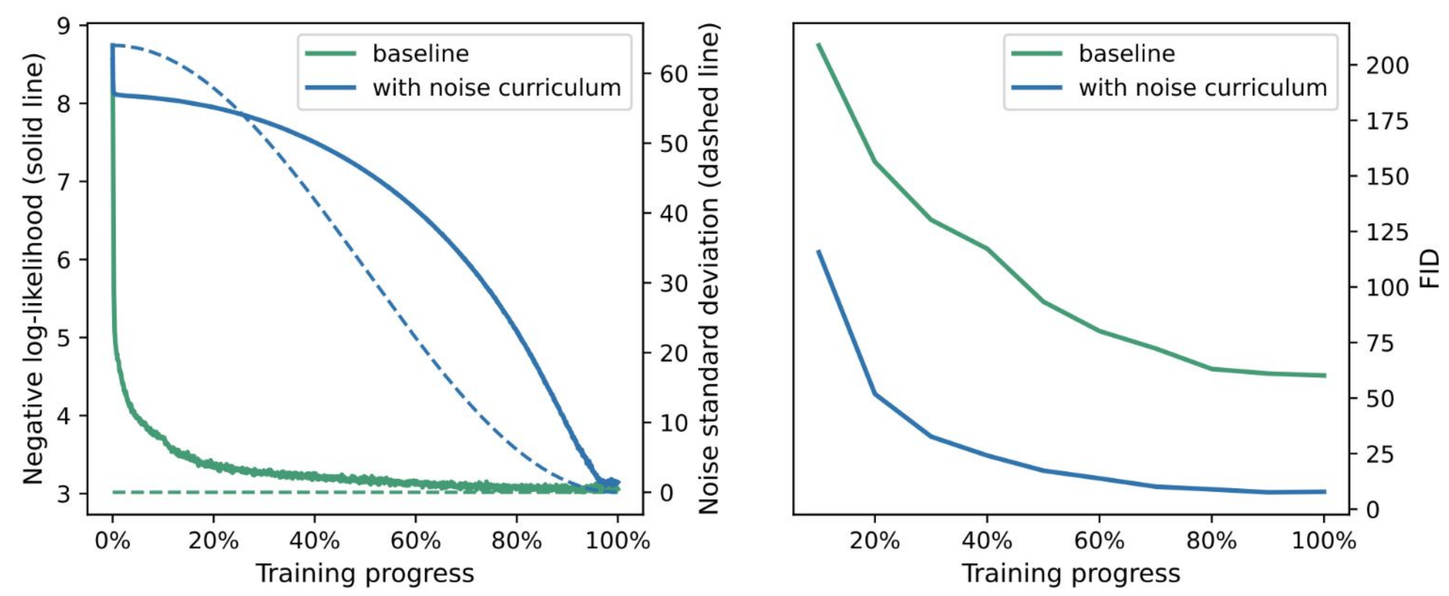
Example 1 Ex. 1 VAE reconstr. Example 2 Ex. 2 VAE reconstr.

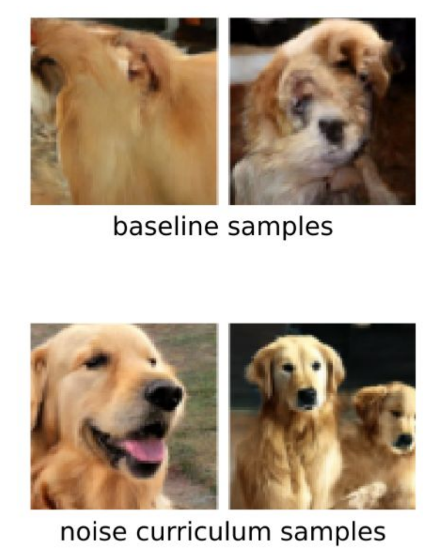
The JetFormer model

- **Challenge & Solution:** Modeling raw pixels autoregressively is computationally costly. JetFormer overcomes limitations of pre-trained tokenizers by combining a normalizing flow (Jet) with a decoder-only transformer, trained end-to-end on raw pixels and text.
- **Core Mechanism:** Jet losslessly encodes images into continuous "soft tokens". The transformer models text tokens and image soft tokens autoregressively, using a GMM loss (à la GIVT) for soft tokens. Jet acts as both encoder (understanding) and decoder (generation).
- **Improving Image Quality:**
  - A **Noise Curriculum** (adding decaying Gaussian noise during training) guides the model to learn high-level visual structure first.
  - Redundancy is handled by **factoring out dimensions** post-flow and modeling them with a Gaussian prior.
  - **Classifier-Free Guidance (CFG).**









	FID	Precision	Recall	NLL
JetFormer-B	7.84	0.75	0.39	3.14
no normalizing flow	117.76	0.17	0.32	6.84
no noise curriculum	44.71	0.45	0.28	3.05
no factored-out dimensions	17.29	0.65	0.26	3.13
no end-to-end training	11.16	0.68	0.33	3.08
learned inv. projection	10.19	0.73	0.32	4.78
no GMM (Gaussian loss only)	9.46	0.77	0.30	3.14
single class token	8.85	0.73	0.37	3.14
PCA preproc. + JetFormer-B	8.79	0.77	0.35	–
PCA preproc. + JetFormer-B (no noise cur.)	13.16	0.71	0.31	–

Ablation of design choices and improvements on ImageNet 256

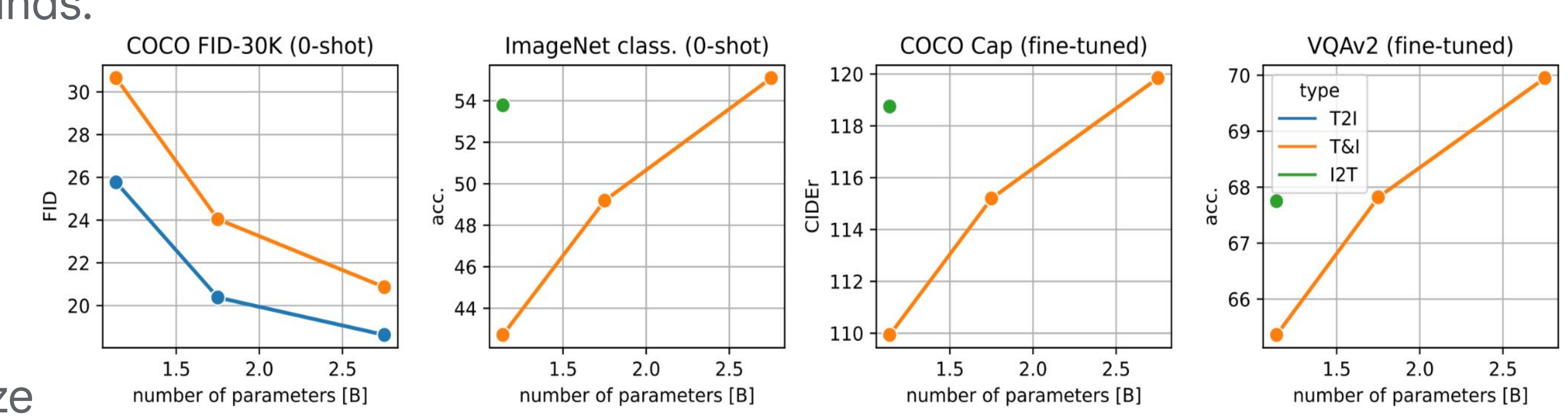
	extra step	COCO cap.	VQAv2
CapPa L/14 (Tschannen et al., 2023)*	–	118.7	68.6
CLIP L/14 (Radford et al., 2021)*	–	118.2	67.9
ARGVLT (T&I) (Kim et al., 2023)	VQ-VAE	94.7	–
MAGVLT Large (T&I) (Kim et al., 2023)	VQ-VAE	110.7	65.7
JetFormer-B (I2T)	–	118.7	67.2
JetFormer-L (T&I)	–	119.8	70.0

Image understanding results (fine-tuned) and comparison with baselines

Key Results

- **Class-conditional image generation (ImageNet 256):**
  - JetFormer achieves **competitive FID (6.64 for L model)** and high recall (0.56), suggesting robustness against mode collapse compared to baselines.
  - Ablations confirm the importance of the normalizing flow, noise curriculum, factoring out dimensions, and end-to-end training.
  - JetFormer is the first model capable of generating high-fidelity images and producing strong log-likelihood bounds.
- **Multimodal generation and understanding (WebLI):**
  - Achieves T2I generation performance competitive with VQ-based models
  - Demonstrates solid I2T understanding (zero-shot classification, fine-tuned captioning/VQA)
  - **Promising scaling trends** as a function of the model size

ImageNet 256 class-conditional samples (CFG=4)



Zero-shot and fine-tuning results for T2I and I2T tasks as a function of model size