Binary Latent Decoder for Text-Conditioned Image Generation

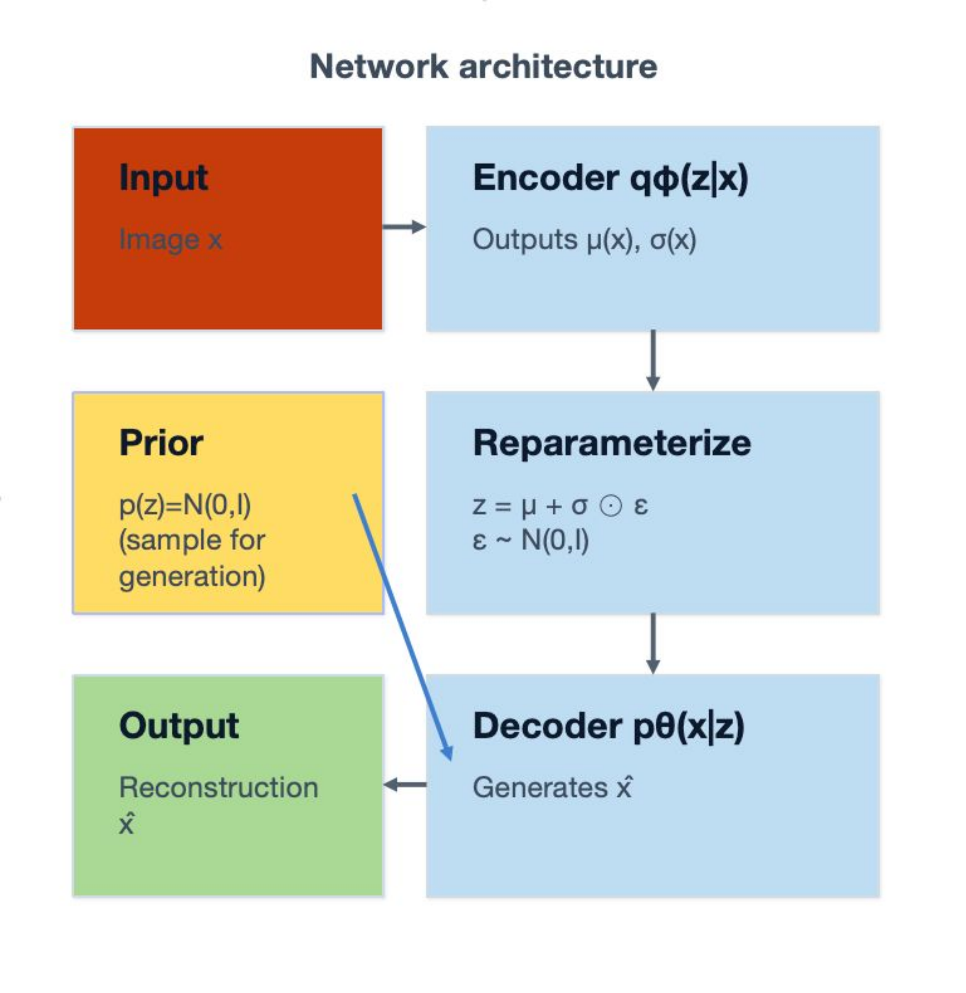
Presenters

- Jince Liu
- Ken Su
- Mihir Sukharamwala

UNIVERSITY OF ALBERTA

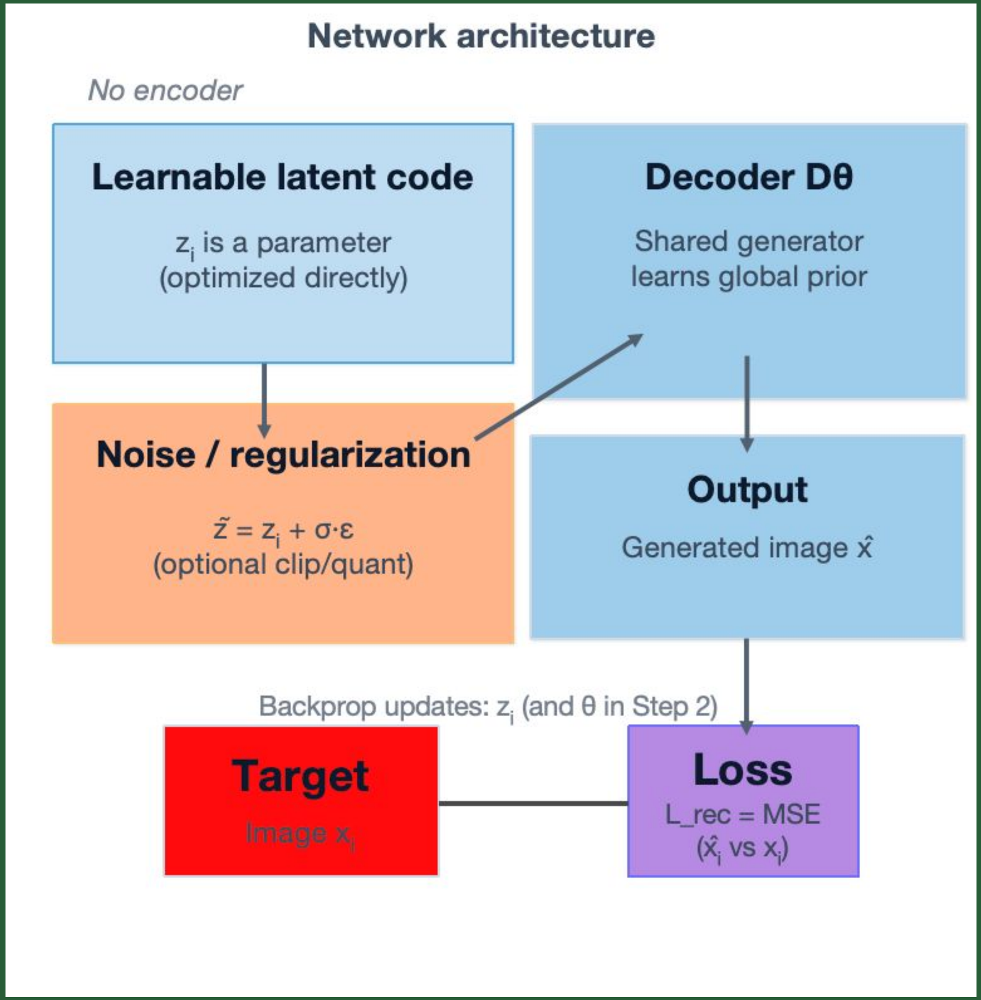
Traditional Variational Autoencoder (VAE)

- Why VAE can generate images: encoder predicts a Gaussian latent distribution ($\mu(x)$, $\sigma(x)$); sample $z = \mu + \sigma \odot \varepsilon$, $\varepsilon \sim N(0,I)$; decode z to an image.
- Text-conditioning is hard: need a strong conditional prior $p(z \mid \text{text})$ and stable alignment so text actually controls generation.
- Heavier training: encoder + decoder → more parameters, more GPU memory/compute, and slower training.



Our Decoder-Only Model (No Encoder)

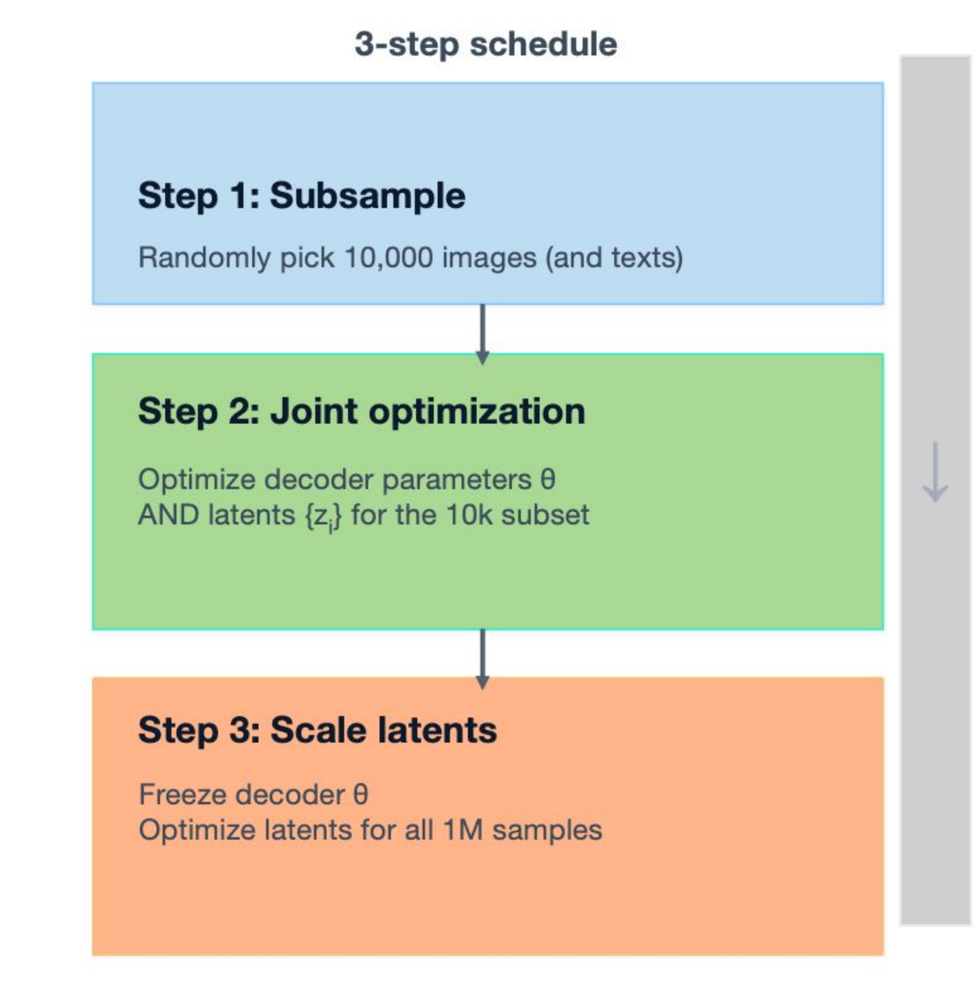
- Decoder only (no encoder) → fewer parameters and faster training.
- Latents are learnable parameters (one latent per sample), not outputs of an encoder.
- VAE stores information mostly in network parameters; our approach stores per-sample information in the latent space, while the decoder learns a shared prior.



UNIVERSITY OF ALBERTA

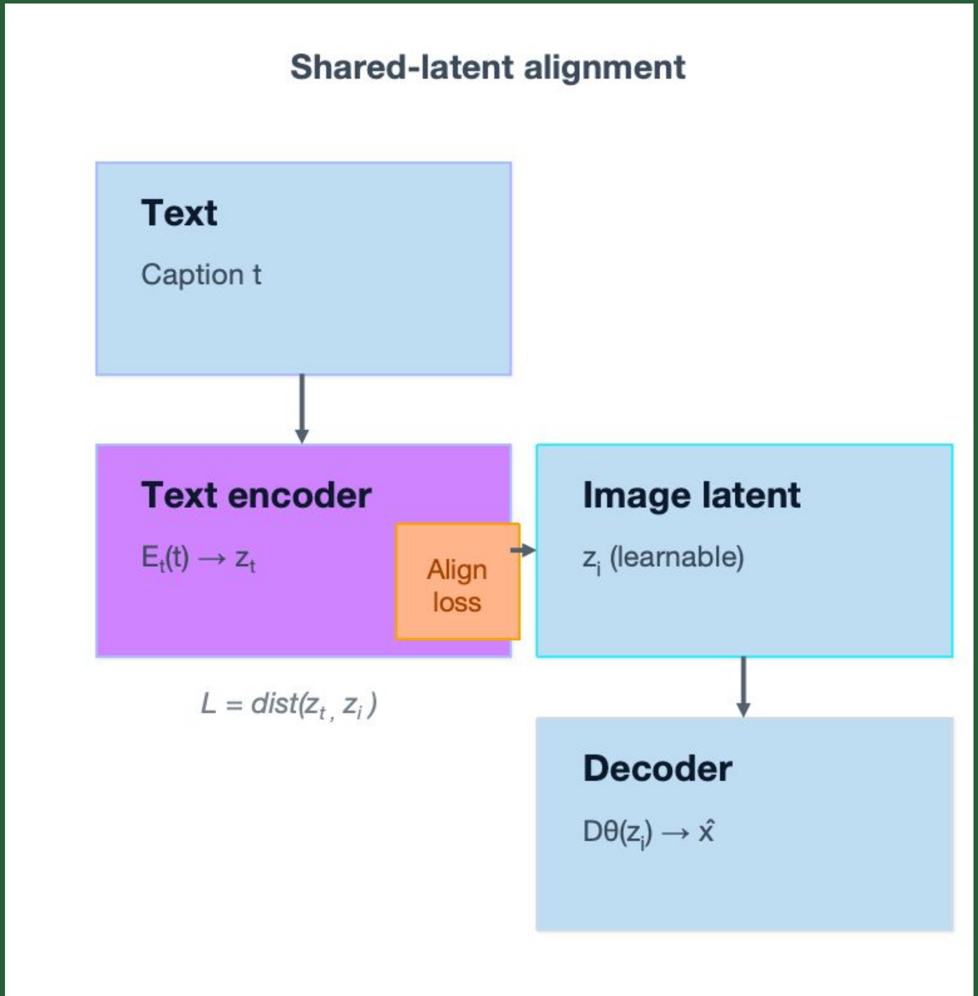
Training Strategy (Limited GPU Resources)

- Full dataset: 1M image–text pairs (GPU constraints prevent full joint training).
- Key idea: train a strong shared decoder first, then adapt per-sample latents at scale.
- Decoder learns a global generation prior; latents later adapt per sample.



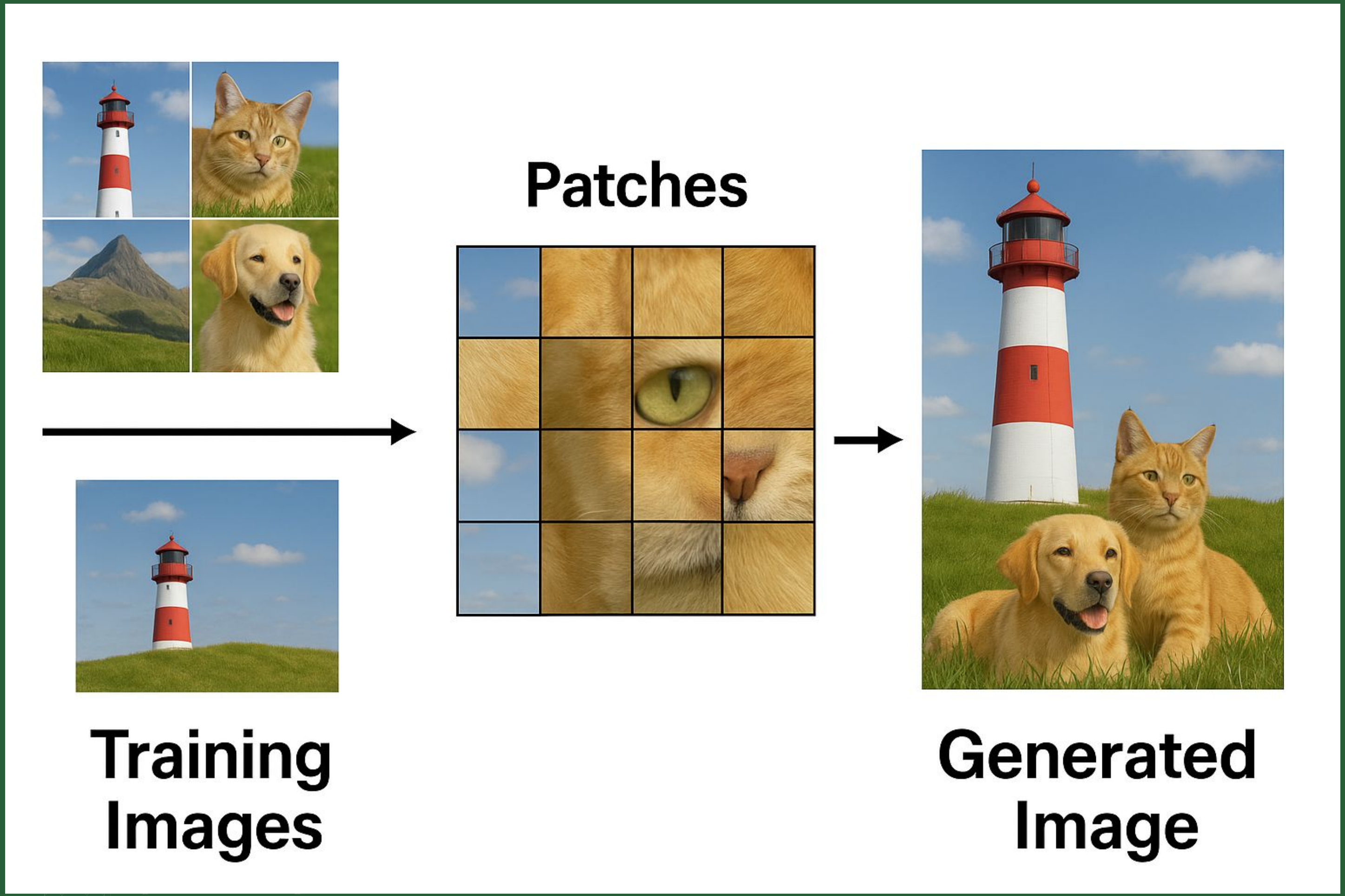
Aligning Text and Images via Shared Latent Space

- We align text and images through a shared latent space.
- Each text latent is linked to its corresponding image latent (paired supervision).
- After alignment, text can retrieve or initialize the image latent used for generation.



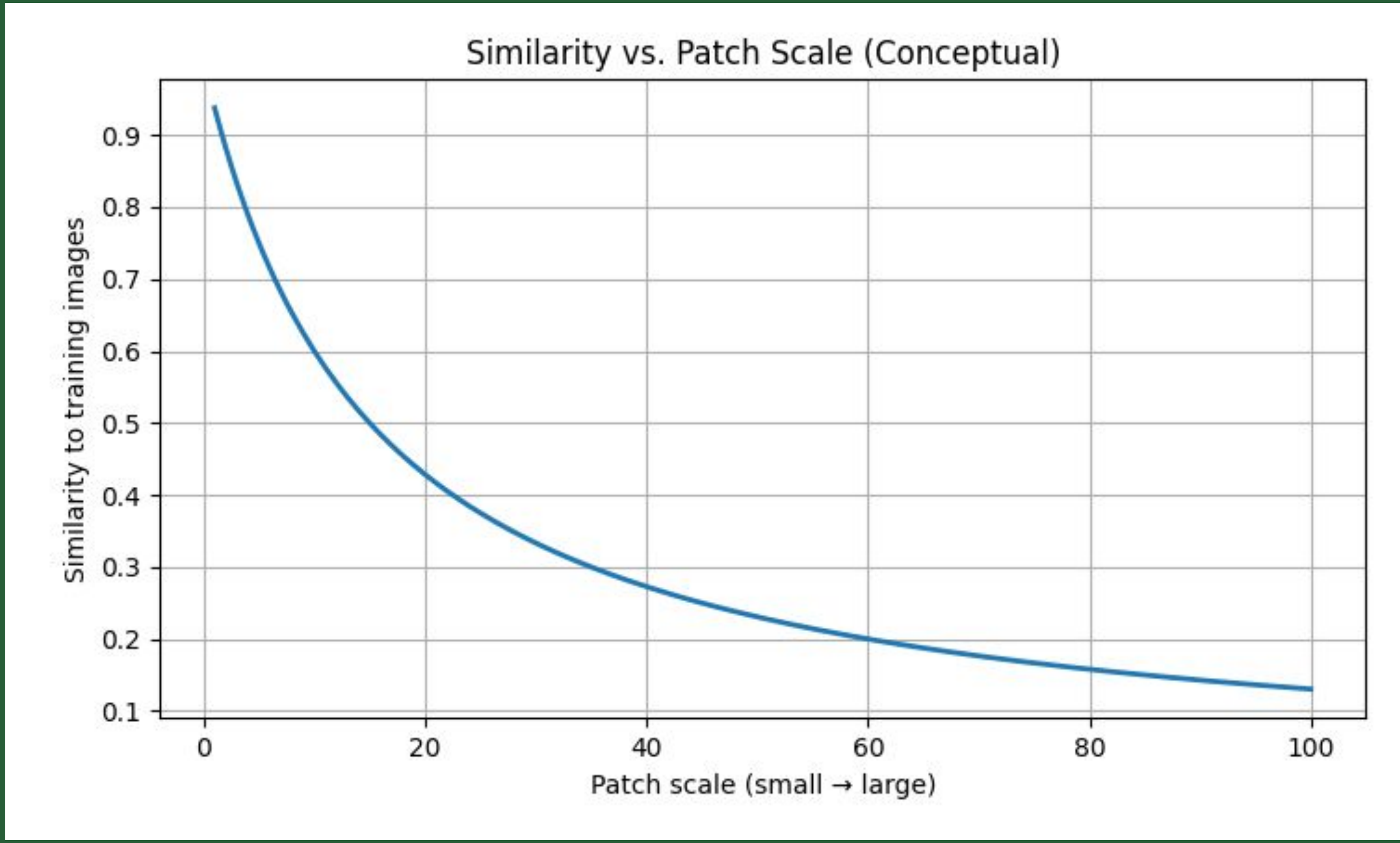
UNIVERSITY OF ALBERTA

What is image generation?



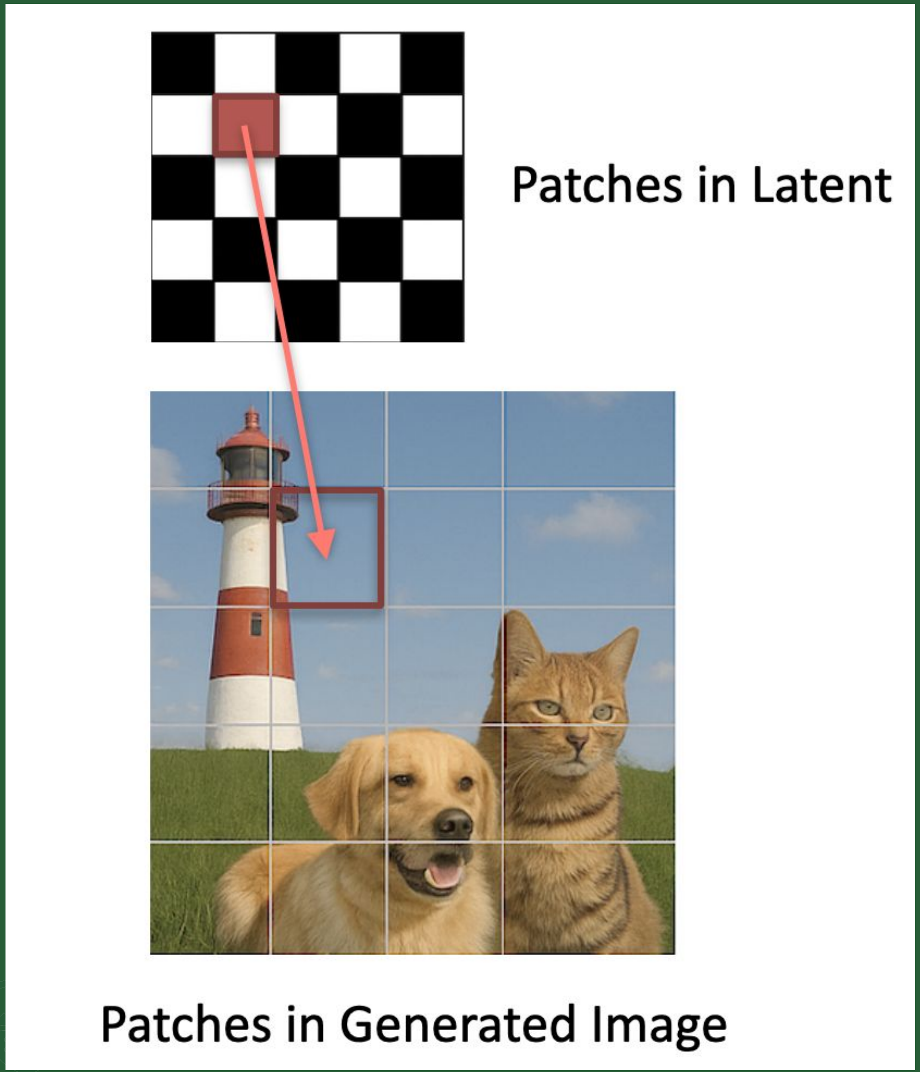
Every patch in generated images should be similar to patches that is already exist in training image.

What is image generation?



As the patch scale increases, the similarity decreases

Locality Constraint in Image Generation



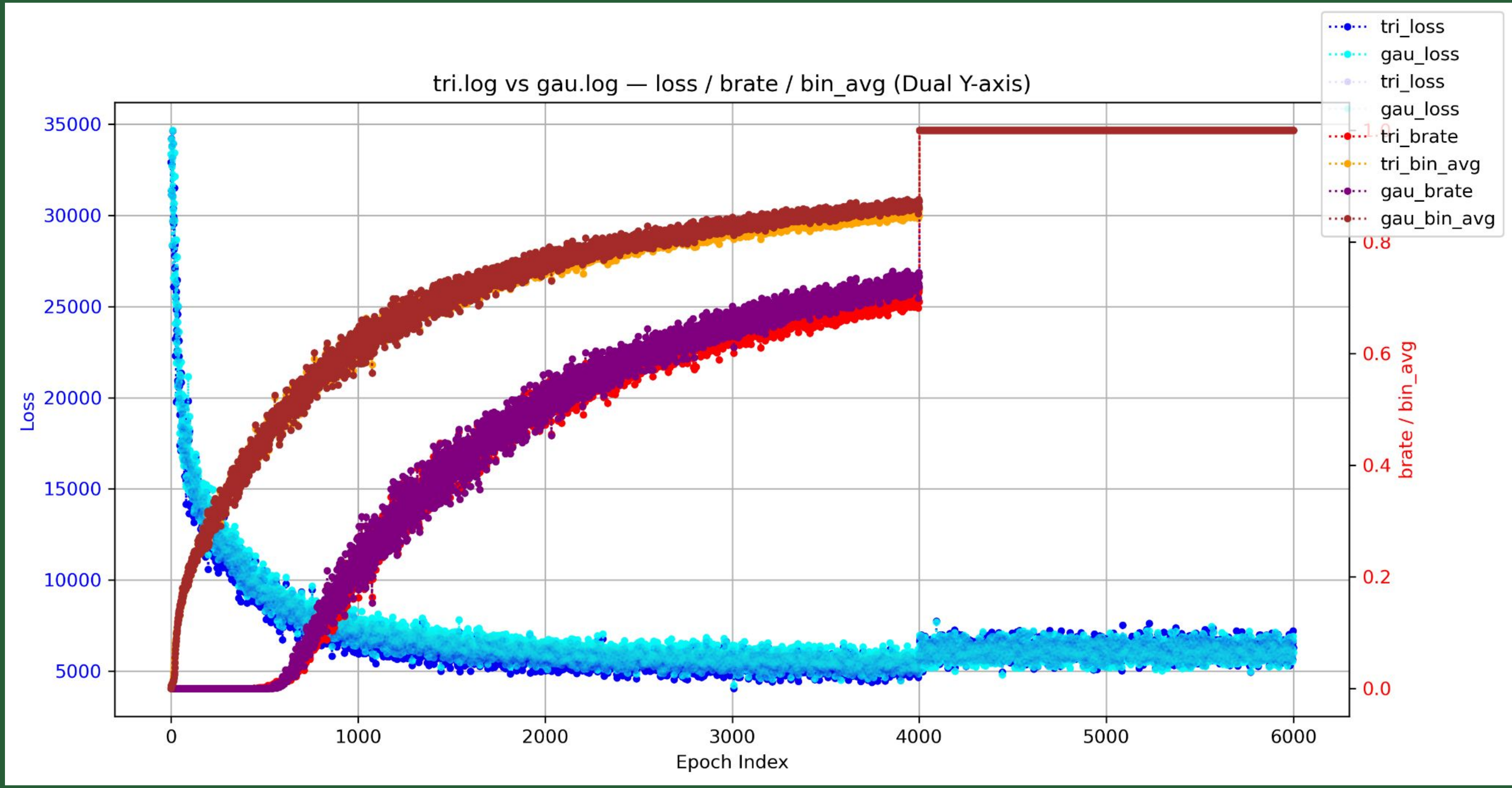
the value of each patch in generated images is determined by a corresponding patch in the latent

Information Capacity and Generalization

- Model capacity should be lower than dataset information
- Excess capacity leads to memorization
- Information is stored in binary latents, not parameters
- Latent size controls generation quality vs generalization

Part 2

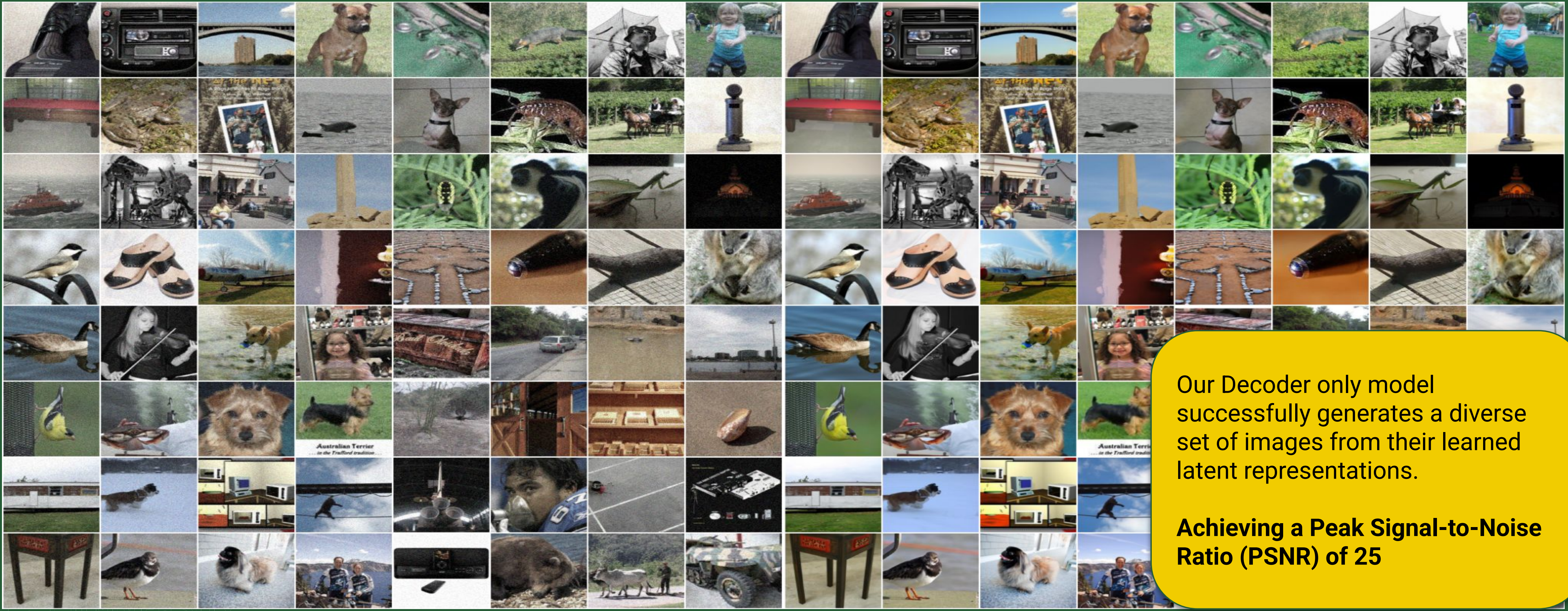
Part 2



tri.log vs gau.log — loss / brate / bin_avg (Dual Y-axis)
tri_loss
gau_loss
tri_loss
gau_loss
tri_brate
tri_bin_avg
gau_brate
gau_bin_avg
Loss
35000
30000
25000
20000
15000
10000
5000
brate / bin_avg
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
Epoch Index

Demonstrating High Fidelity Image Generation

Generated Images — Original Dataset Images



Our Decoder only model successfully generates a diverse set of images from their learned latent representations.

Achieving a Peak Signal-to-Noise Ratio (PSNR) of 25

The Generation Pipeline: From Text to Image

The shared latent space acts as a **bridge between modalities**, allowing us to translate semantic information from text into a visual representation.

A Case Study: “A dog laying on the floor”

Input

“A Dog Laying on the floor”

Latent Space

Text Latent:
0101…..01010

↓

Image Latent:
1010…..1010

Output

Core Research and Contribution

Quantifiable Information Capacity

Binary latent representations allow the amount of stored information to be explicitly quantified, a clarity not possible with continuous latents or model parameters.

Architectural Simplicity & Efficiency

A simple decoder-only architecture without an encoder streamlines the training process and significantly reduces GPU resource requirements.

Novel Optimization Strategy

Our experiments validate the feasibility of jointly optimizing latent codes and decoder parameters, a key enabler for our architecture.

Generation Stability

Our experiments validate the feasibility of jointly optimizing latent codes and decoder parameters, a key enabler for our architecture.

An Honest Assessment: Current Limitations

Visual Fidelity

The quality of generated images, while promising, requires further improvement to reach state-of-the-art levels.

Text-Image Coherence

A stronger and more nuanced alignment between the semantics of the input text and the content of the visual output is needed.

Future Enhancements

1.

Scale the Foundation: Full Dataset Training

Action: Secure additional GPU resources to train the model on our complete I-million-pair dataset.

Goal: To allow the decoder to learn a more robust and learn a more robust and comprehensive global generation prior.

2.

Enhance the Engine: Improve Image Quality

Action: Investigate enhanced decoder architectures and explore the potential of multi-stage generation.

Goal: To directly address the current limitations in visual fidelity.

3.

Refine the Bridge: Strengthen Text-Image Alignment

Action: Develop and integrate more effective alignment methods between text and image latent spaces.

Goal: To ensure more accurate and efficient semantic translation for conditional generation.

Q&A

UNIVERSITY OF ALBERTA