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[†]

[μ] Meta
[δ] Waymo [σ] University of Southern California

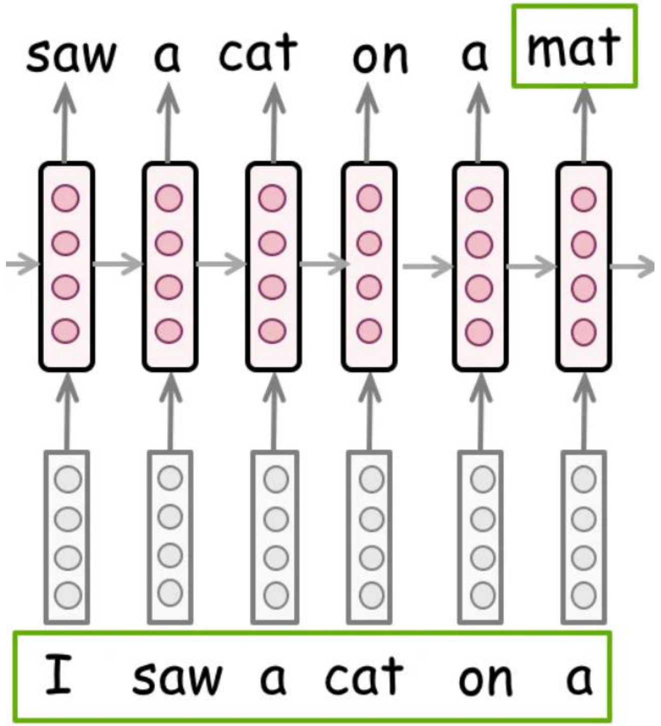
PARIS 2024
PARIS 2024
2024
PARIS 2024
TÜRKİYE

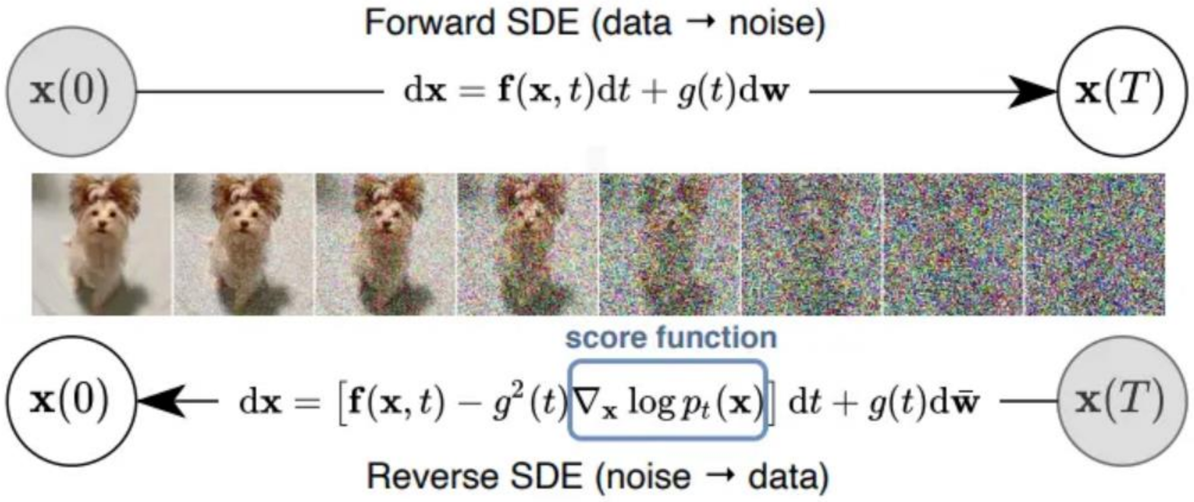
Hybrid Architecture

Transformer vs Diffusion

- Transformer
  - Next token prediction (autoregressive)
  - One network pass for each token
  - Good for language generation

- Diffusion
  - Sample-level prediction
  - Iterative for each sample
  - Good for visual generation




Forward SDE (data → noise)
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
$\mathbf{x}(T)$
score function
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$
$\mathbf{x}(T)$
Reverse SDE (noise → data)

https://hyperskill.org/learn/step/24084

https://www.superannotate.com/blog/diffusion-models

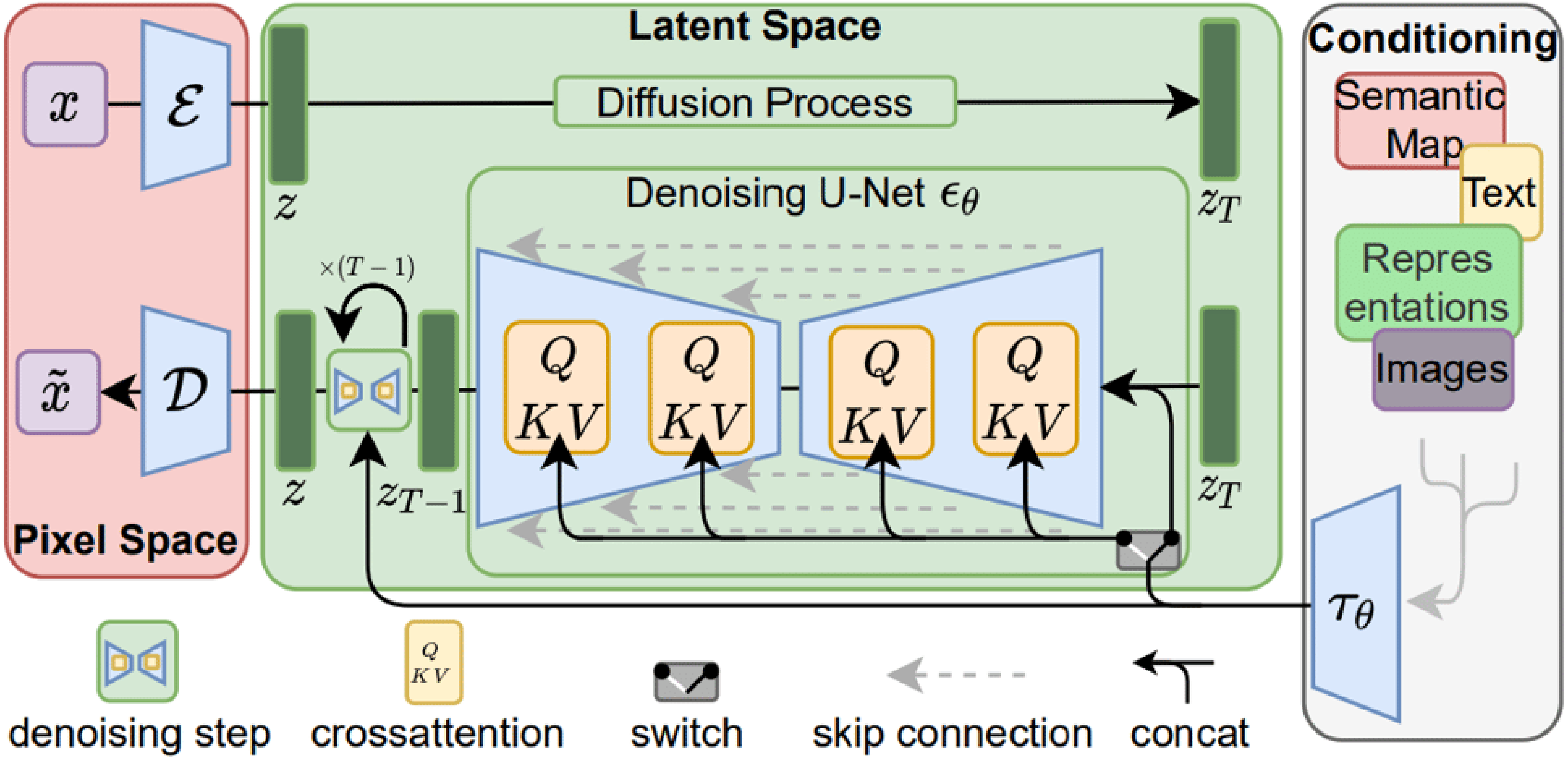
Hybrid Modality

Language vs Visual

- Language
  - Discrete (from vocabulary)
  - Cross-entropy loss
  - **2** ("elephant") *is not* close to **3** ("car")
- Visual
  - Continuous
  - L1 loss
  - **2** *is* close to **3**

Hybrid Visual Generation

Diffusion: Latent Diffusion Model (Stable Diffusion)

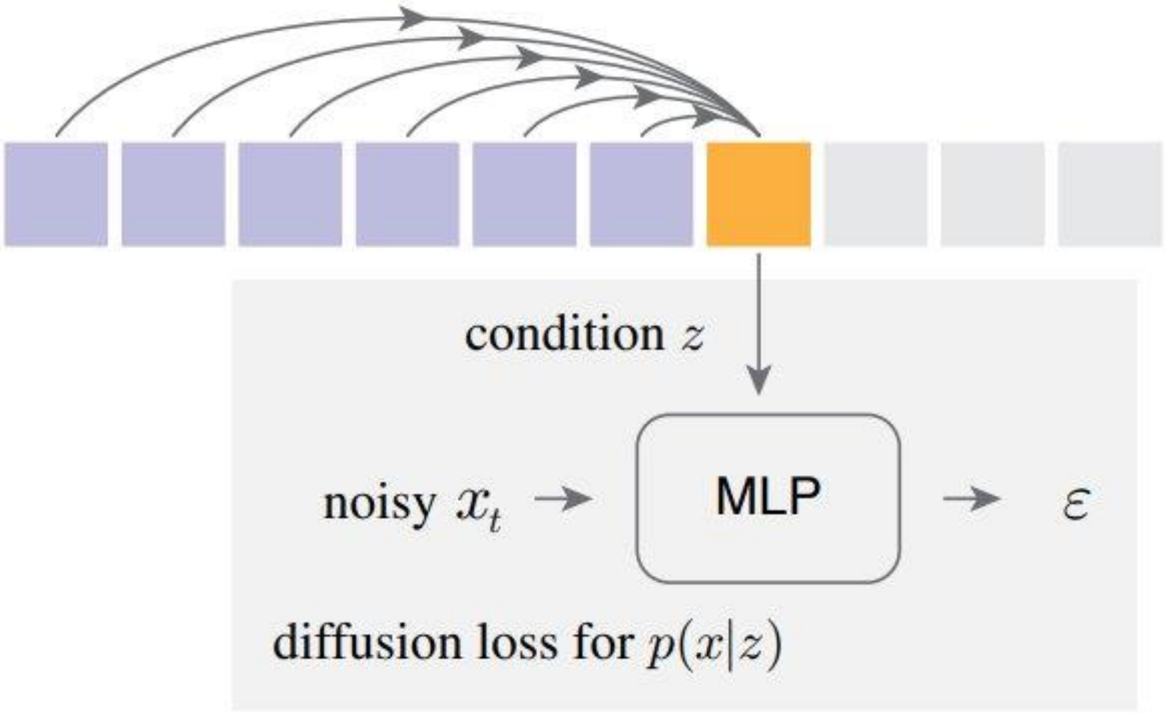


High-Resolution Image Synthesis with Latent Diffusion Models

Hybrid Architectural Visual Generation

Autoregressive: Masked Autoregressive (MAR)

- Continuous Autoregressive
- Single modality

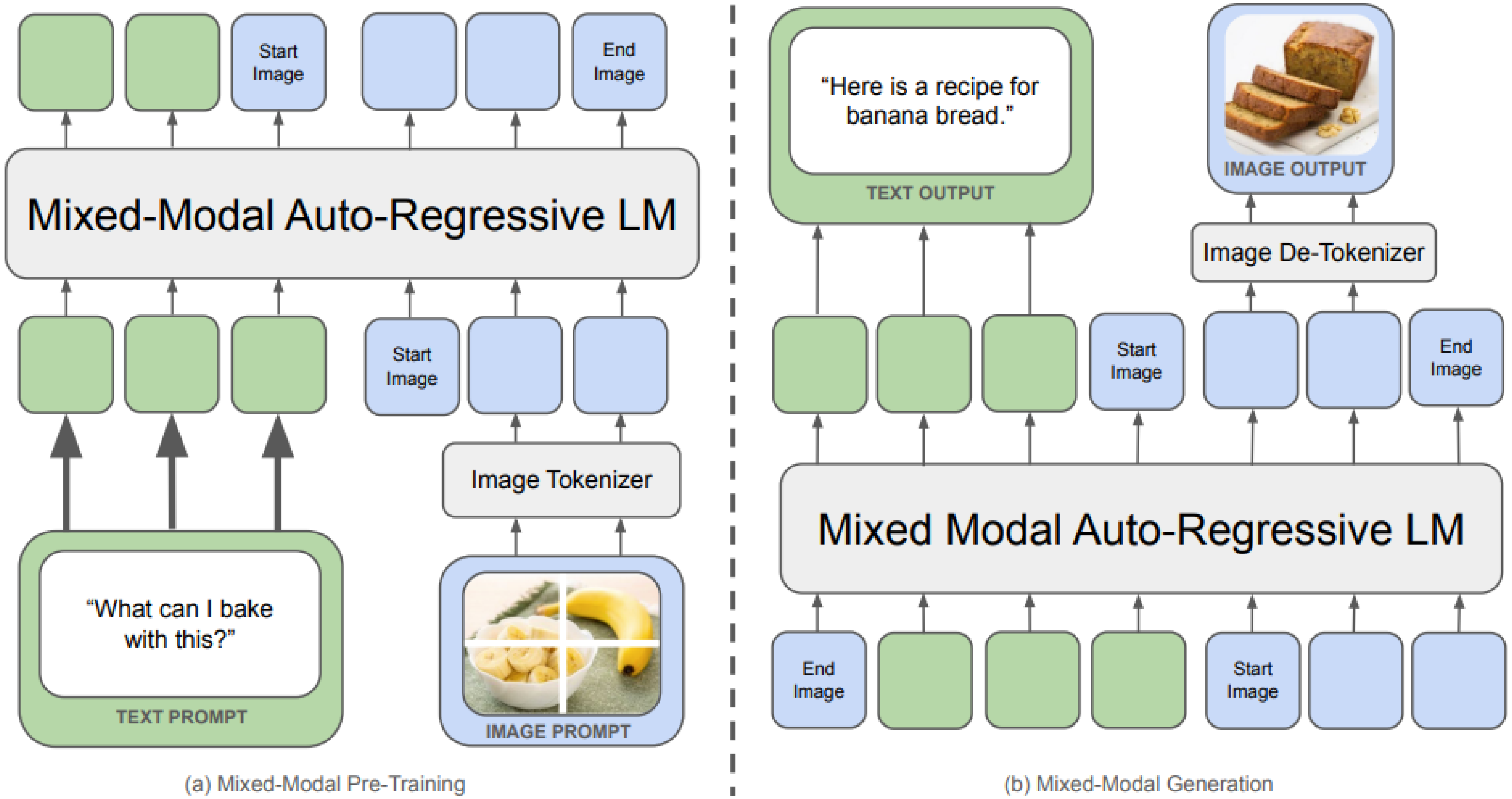


Autoregressive Image Generation without Vector Quantization

Hybrid Modality Language Generation (MLLM)

Autoregressive: GPT-4o (Native), LLaVA, Chameleon

- Single Architecture



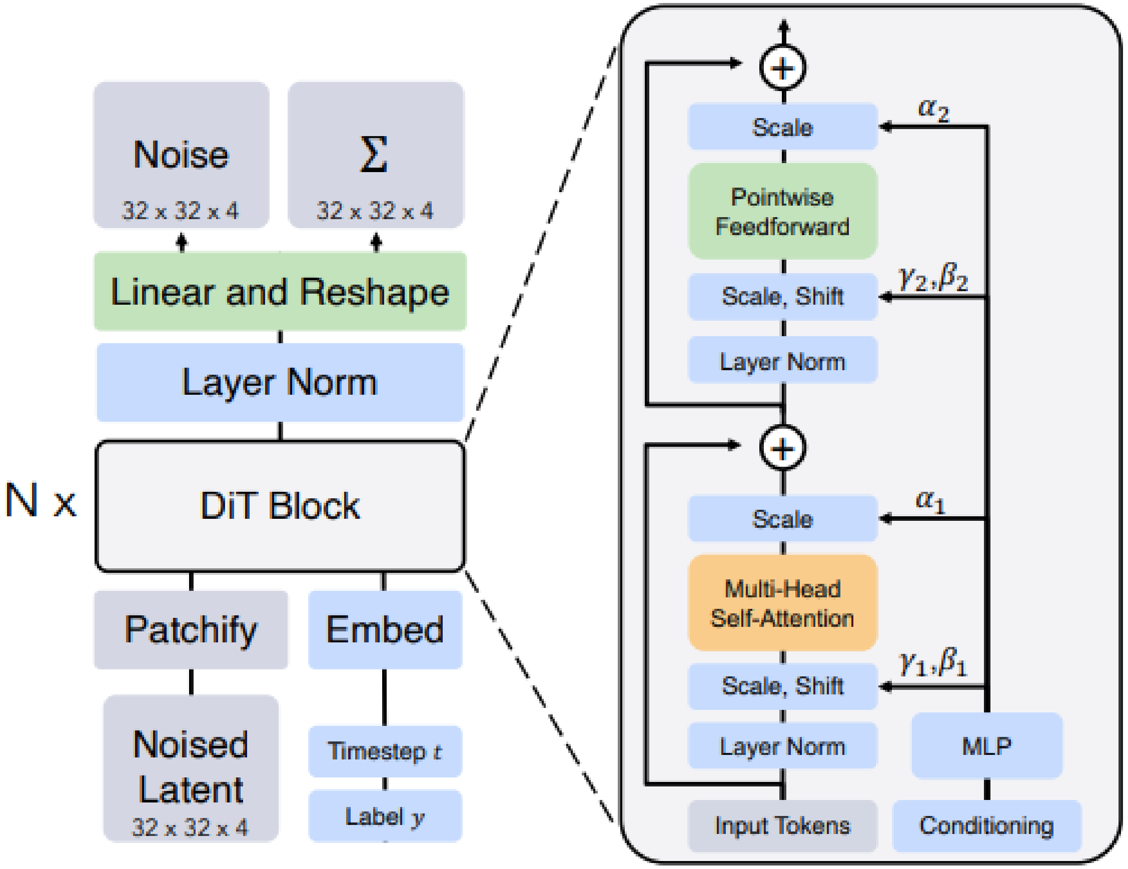
(a) Mixed-Modal Pre-Training

(b) Mixed-Modal Generation

Chameleon Team: Mixed-modal early-fusion foundation models

Building Block: Diffusion Transformer (DiT)

- Transformer replaces U-Net in Diffusion.



Latent Diffusion Transformer

DiT Block with adaLN-Zero

Scalable Diffusion Models with Transformers

Hybrid Generation: Transfusion: Overview

- Autoregressive (a single Transformer to do everything).
- Diffusion on the image part.

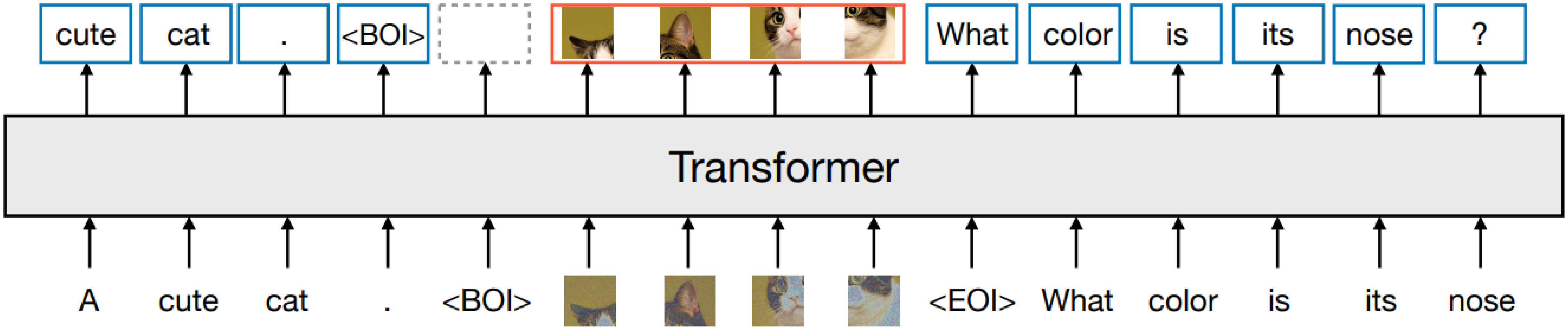
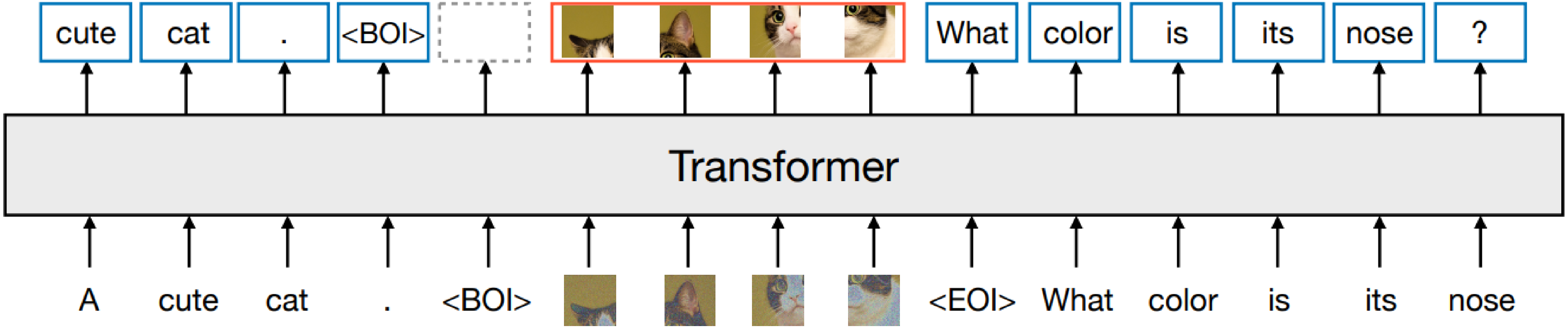


Figure 1: A high-level illustration of Transfusion. A single transformer perceives, processes, and produces data of every modality. Discrete (text) tokens are processed autoregressively and trained on the next token prediction objective. Continuous (image) vectors are processed together in parallel and trained on the diffusion objective. Marker BOI and EOI tokens separate the modalities.

Transfusion: Mechanism

- **Testing**
  - LM mode: Transformer as a NTP Language Model to sample text until <BOI>.

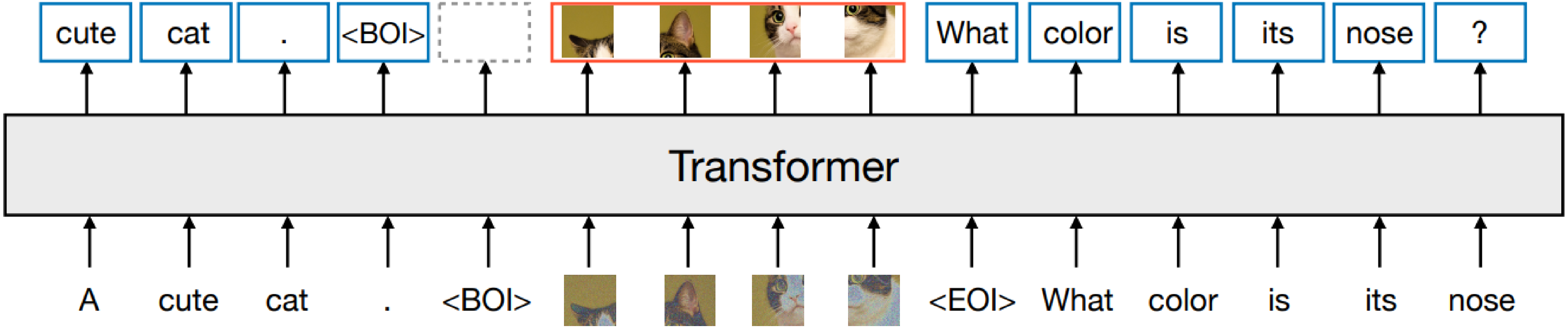
    Switch To:
  - Diffusion mode: Transformer as in a latent DiT to sample the whole image (with desired size). Append <EOI> and switch back to LM mode.
    - Previous textual tokens are input along with the latent image tokens, while noises are added only to the image tokens.



Transfusion: Mechanism

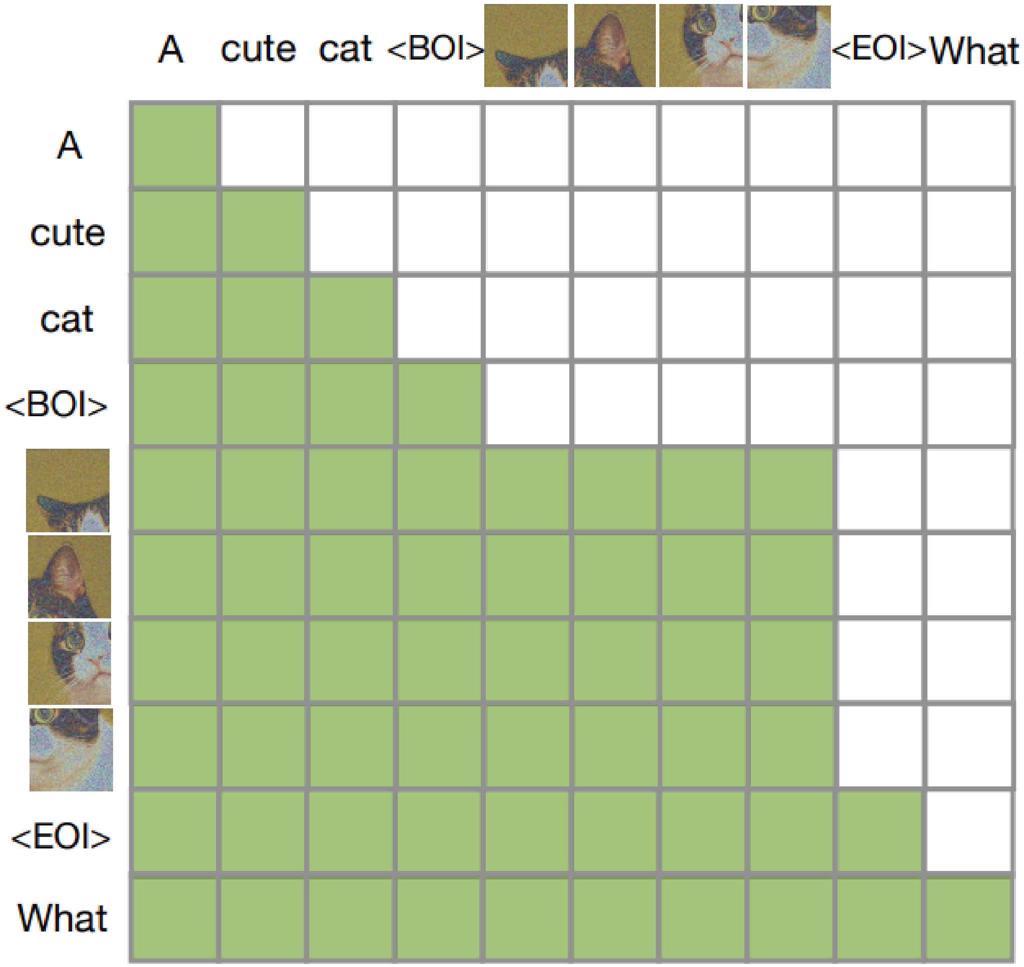
$$\mathcal{L}_{\text{Transfusion}} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{DDPM}}$$

- **Training**
  - Language Modeling on Transformer
    - Enable textual generation.
    - Enable the multimodal tokens to obtain context from each other, which is also useful for image generation.
  - Diffusion Modeling on Transformer
    - Enable the image tokens to be iteratively refined for image generation.

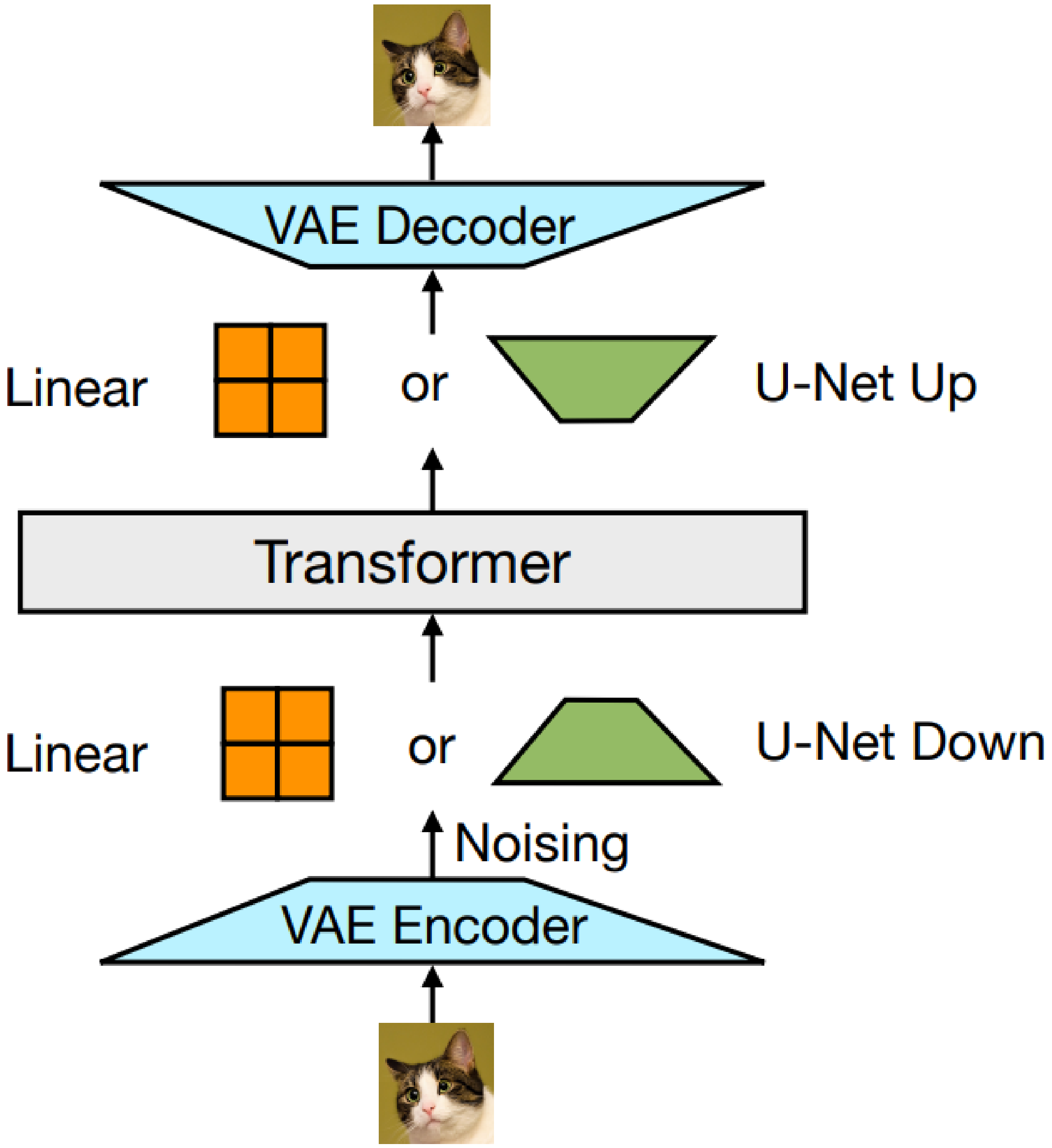


Transfusion: Attention Mask

- In either mode, each token can attend to both textual and image tokens.
- Textual tokens are causal, while image tokens are not.

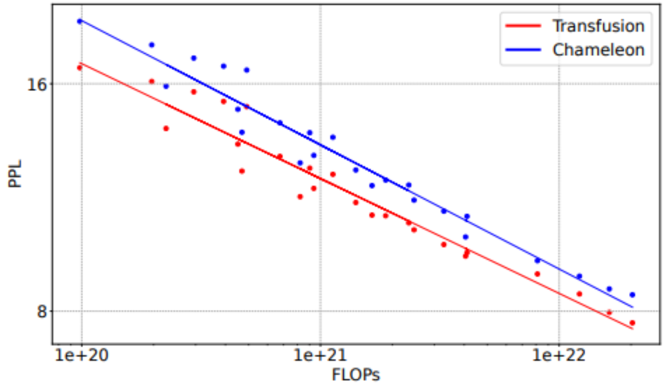


Transfusion: from Image to Transformer features

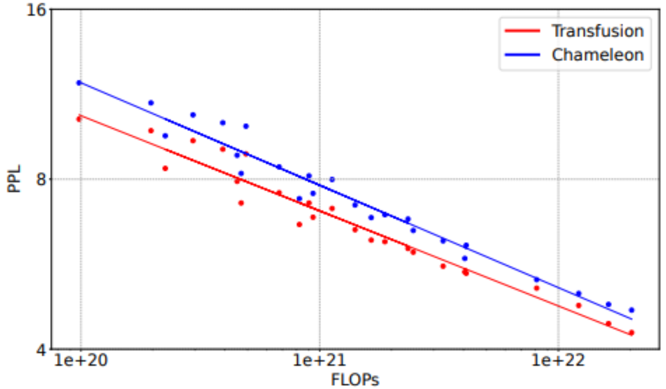


Experiments

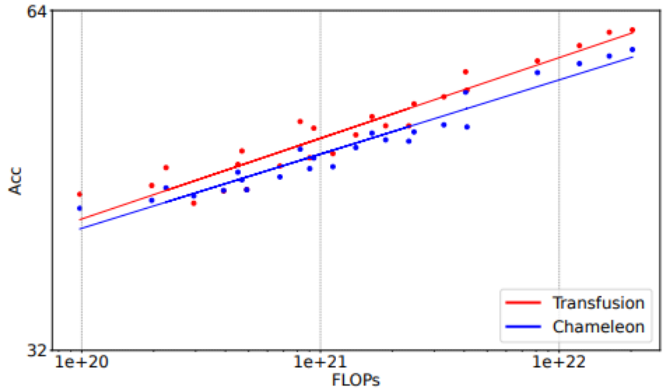
- Scales better than fully autoregressive Chameleon (also by Meta, earlier 2024)
- Perplexity.
- Image to image?
- Image understanding?



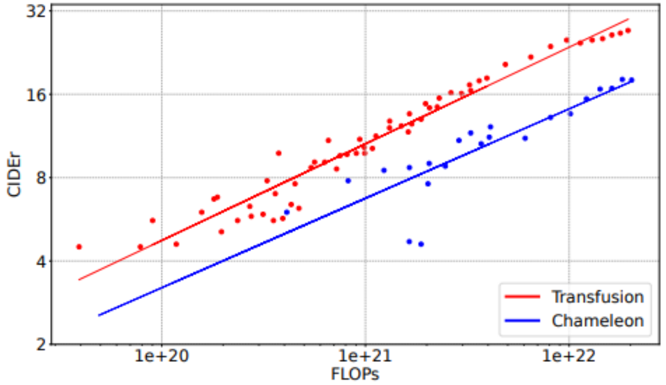
C4 Perplexity



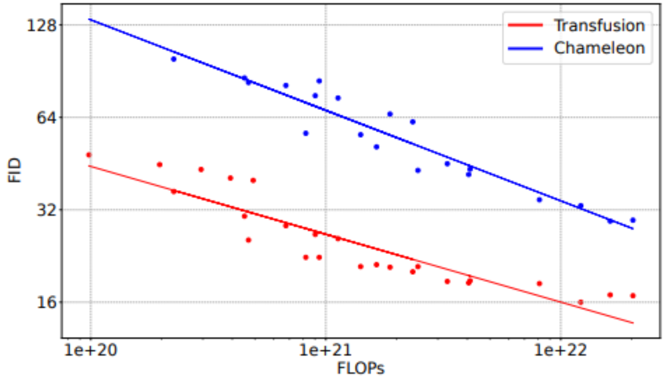
Wikipedia Perplexity



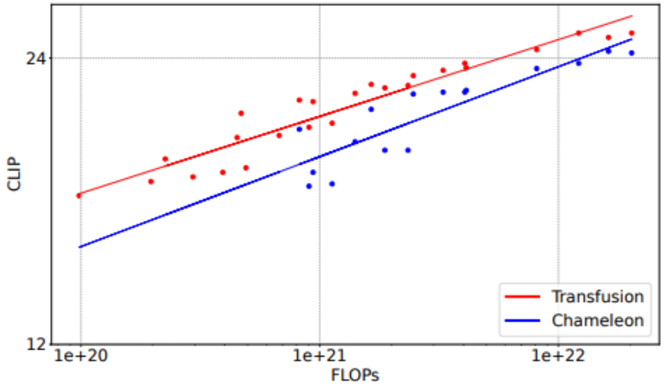
Llama 2 Eval Suite Accuracy



MS-COCO 5k CIDEr



MS-COCO 30k FID



MS-COCO 30k CLIP

Downtown Seattle at sunrise. detailed ink wash.

A car made out of vegetables.

A sign that says “Diffusion".

A black basketball shoe with a lightning bolt on it.

an espresso machine that makes coffee from human souls, high-contrast painting.

Intricate origami of a fox and a unicorn in a snowy forest.

a yellow wall with two framed sketches

A crab made of cheese on a plate.

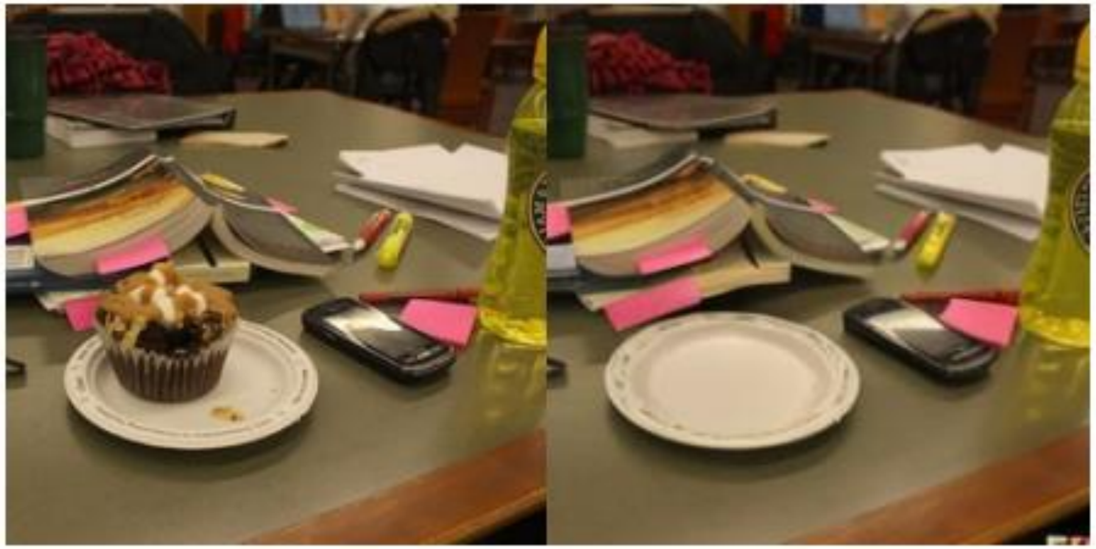
A single beam of light enter the room from the ceiling. The beam of light is illuminating an easel. On the easel there is a Rembrandt painting of a raccoon.

White Cycladic houses with blue accents and vibrant magenta bougainvillea in a serene Greek island setting.

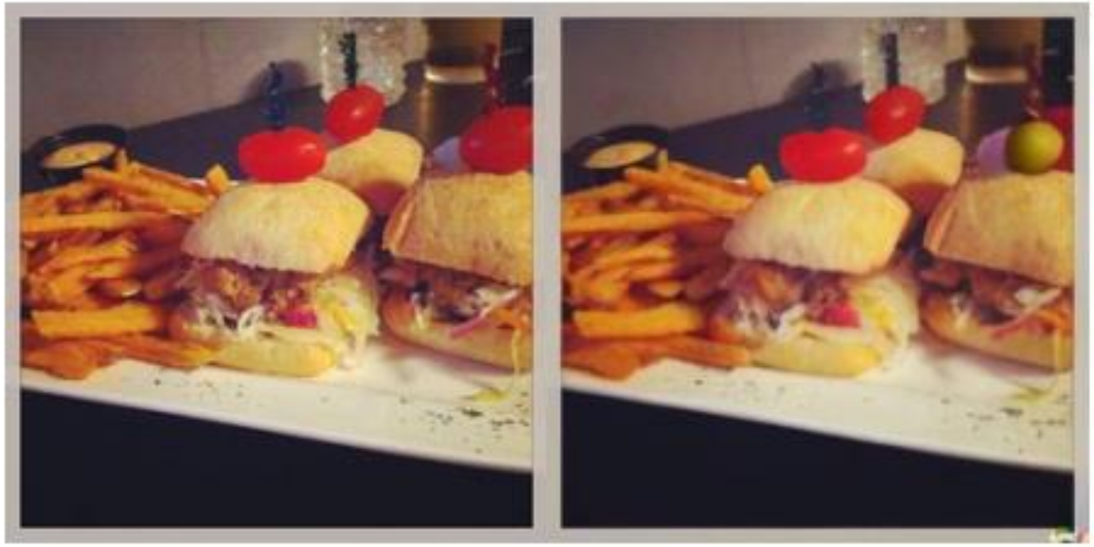
BE
EXCELLENT
TO EACCH
OTTOHER

The saying “BE EXCELLENT TO EACH OTHER" written in a stained glass window.

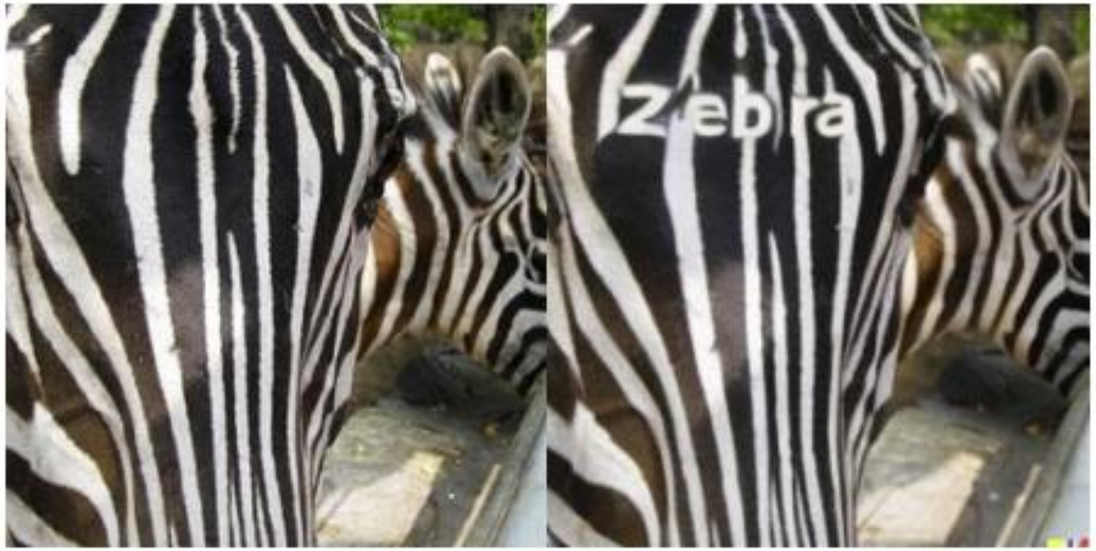
dark high contrast render of a psychedelic tree of life illuminating dust in a mystical cave.



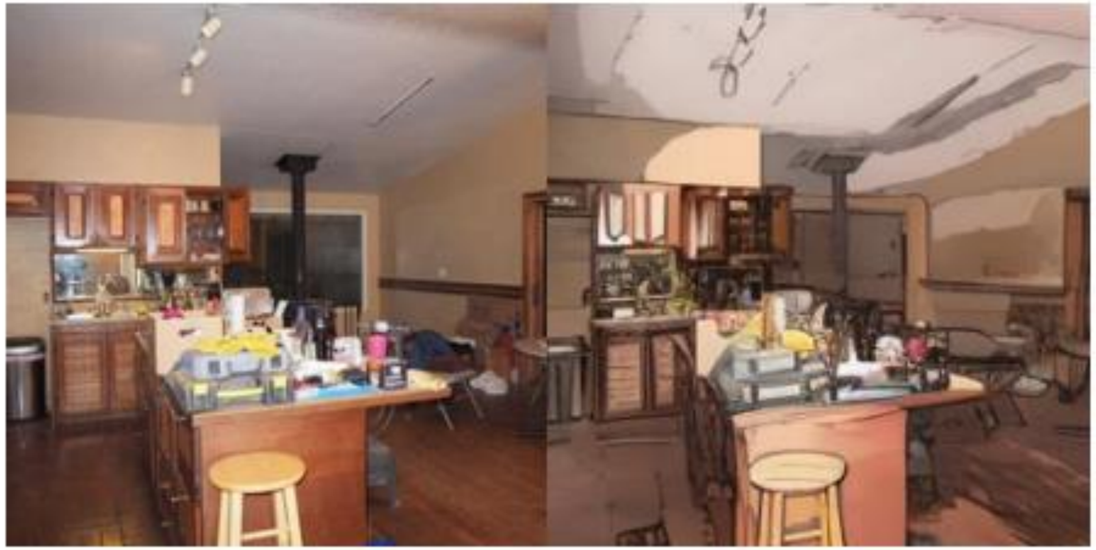
Remove the cupcake on the plate.



Change the tomato on the right to a green olive.



Write the word "Zebra" in Arial bold.



Change this to cartoon style.

Questions?