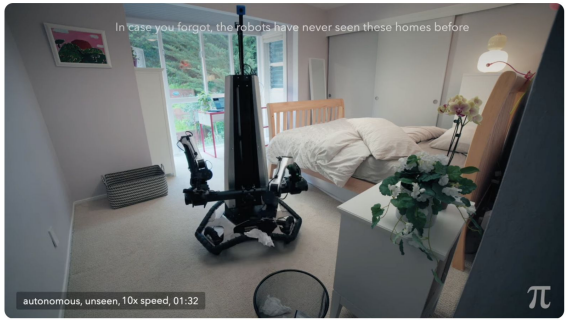
# **Transfusion:** Predict the Next Token and Diffuse Images with One Multi-Modal Model

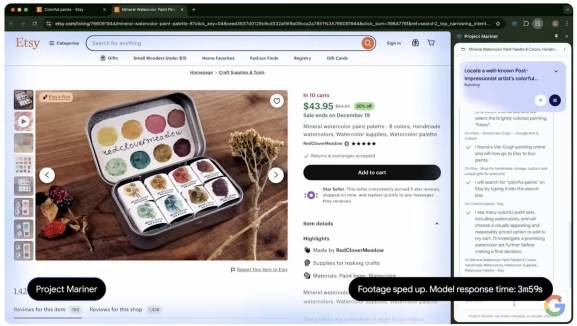
*Chunting Zhou*,* ***Lili Yu*****, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, Omer Levy*


ICLR 2025

# Multimodal AI getting Ubiquitous




Physical Intelligence $\pi_{0.5}$

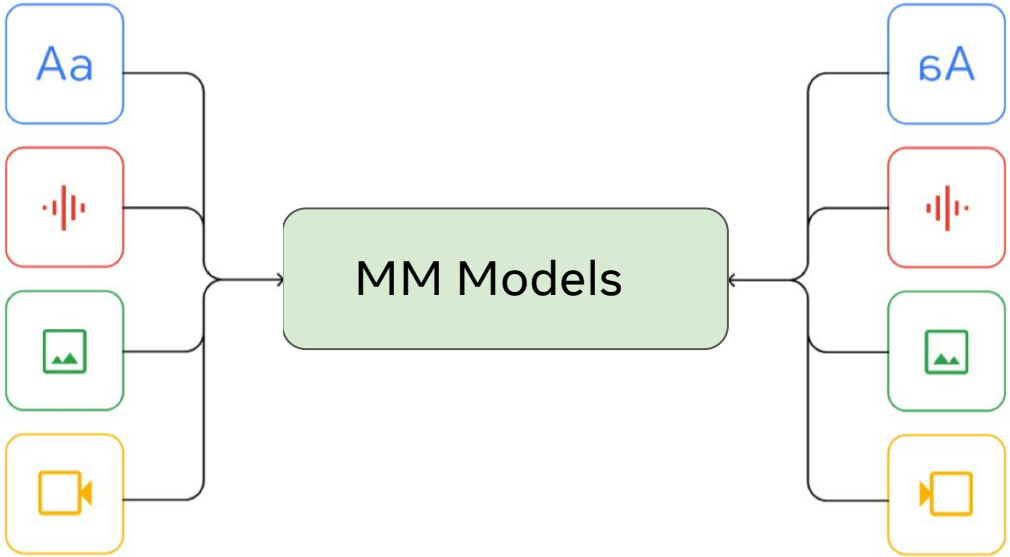



Introducing Gemini 2.0 | Our most capable AI model yet

Gemini 2.0 Multimodal Agent

OpenAI Sora, text-2-video model

# Any-to-any Generative Pretraining

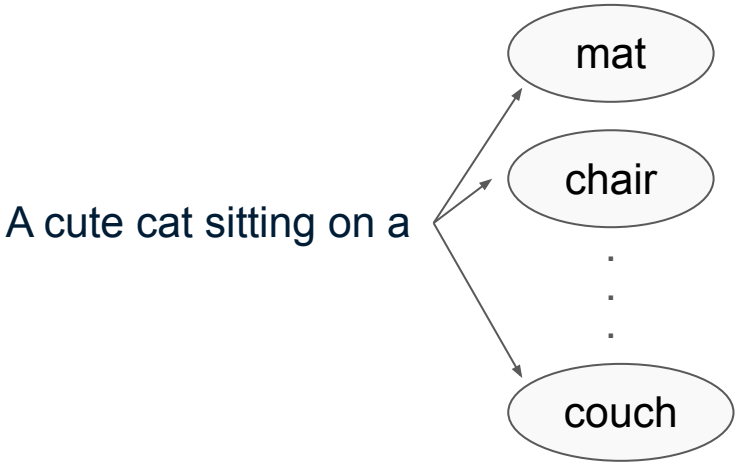



**A single simple model for all modalities, that can perceive, process, and produce both discrete and continuous elements.**

# The problem: no unified model

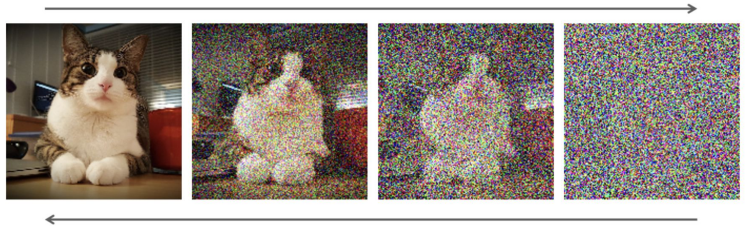
SotA text generation model

Predict the next token




SotA image generation model

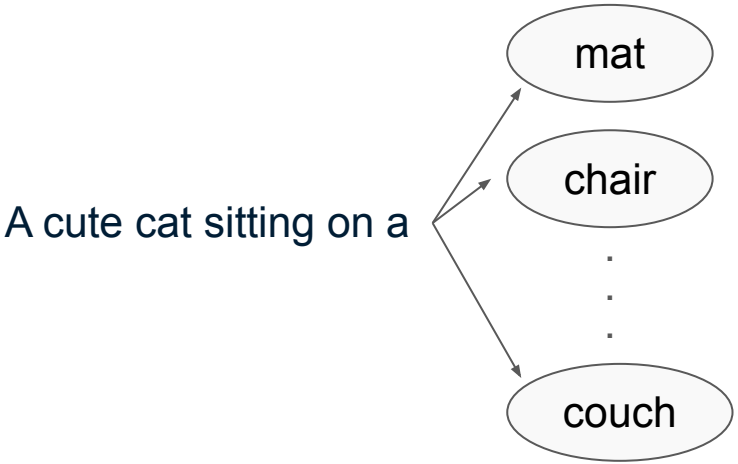
Noising + denoising



# The problem: no unified model

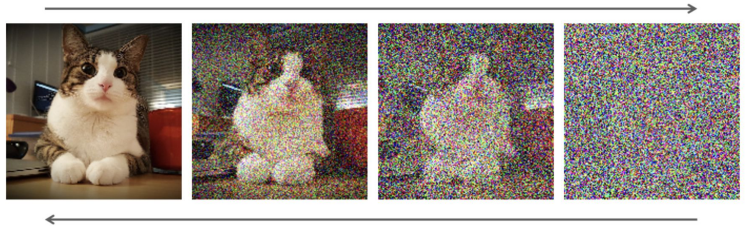
SotA text generation model

Tokenization: discrete tokens
Loss: cross-entropy loss
Architecture: causal transformer




SotA image generation model

Tokenization: continuous vector
Loss: diffusion/flow loss
Architecture: bidirectional transformer



- How to train a single model to generate SOTA text and image?
- How does transfusion perform on text and image generation?
- Does transfusion generalize to new modality combinations?

How to train a single model to generate SOTA text and image?

# Problem: no unified model

SotA text generation model

Tokenization: discrete tokens
Loss: cross-entropy loss
Architecture: causal transformer

SotA image generation model

Tokenization: continuous vector
Loss: diffusion/flow loss
Architecture: bidirectional transformer

# Transfusion: best of both worlds

SotA text generation model

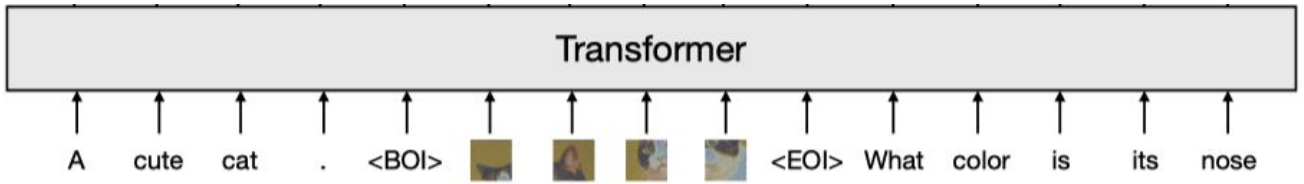
Tokenization: discrete tokens
Loss: cross-entropy loss
Architecture: causal transformer

SotA image generation model

Tokenization: continuous vector
Loss: diffusion/flow loss
Architecture: bidirectional transformer

# Transfusion:

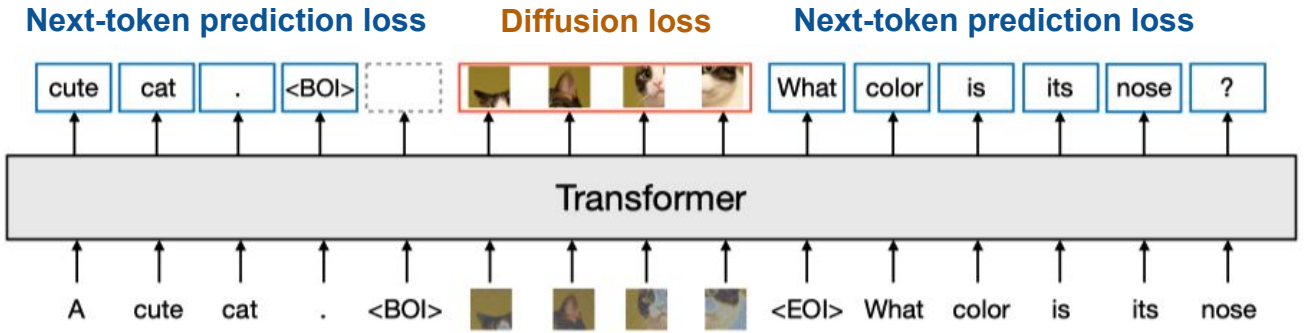
| SotA text generation model | SotA image generation model |
| --- | --- |
| Tokenization: discrete tokens | Tokenization: continuous vector |
| Loss: cross-entropy loss | Loss: diffusion/flow loss |
| Architecture: causal transformer | Architecture: bidirectional transformer |




Data is any arbitrary sequence of interleaved text and images

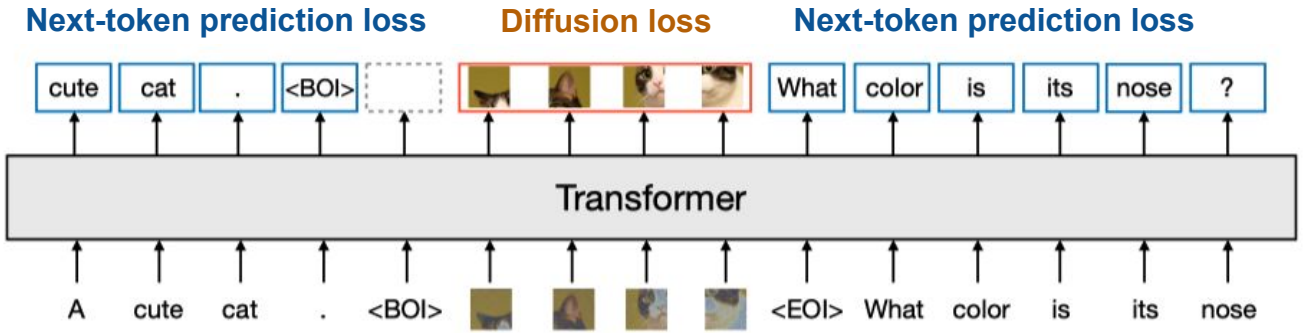
# Transfusion:

| SotA text generation model | SotA image generation model |
| --- | --- |
| Tokenization: discrete tokens | Tokenization: continuous vector |
| Loss: cross-entropy loss | Loss: diffusion/flow loss |
| Architecture: causal transformer | Architecture: bidirectional transformer |



# Transfusion:

| SotA text generation model | SotA image generation model |
| --- | --- |
| Tokenization: discrete tokens | Tokenization: continuous vector |
| Loss: cross-entropy loss | Loss: diffusion/flow loss |
| Architecture: causal transformer | Architecture: bidirectional transformer |




$$\mathcal{L}_{\text{Transfusion}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{Diffusion}}$$

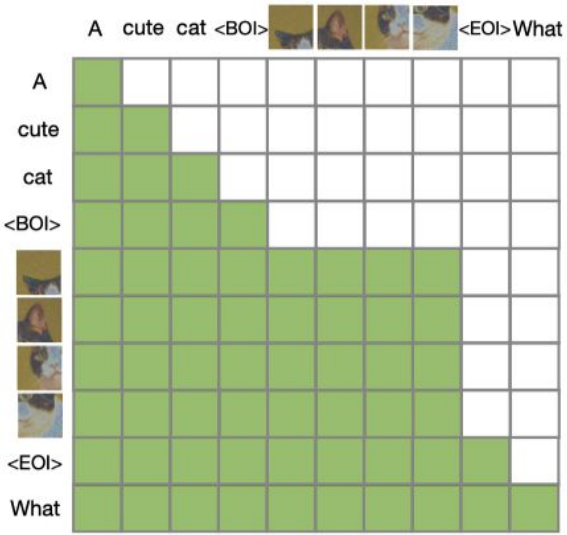
# Transfusion:

SotA text generation model

Tokenization: discrete tokens
Loss: cross-entropy loss
Architecture: causal transformer

SotA image generation model

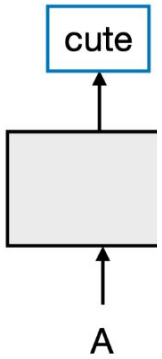
Tokenization: continuous vector
Loss: diffusion/flow loss
Architecture: bidirectional transformer



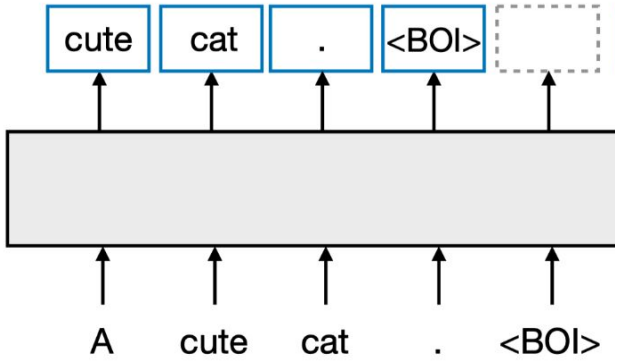

Transfusion uses global causal and image bidirectional attention.

Dramatically improves image generation, and is critical for scaling up.

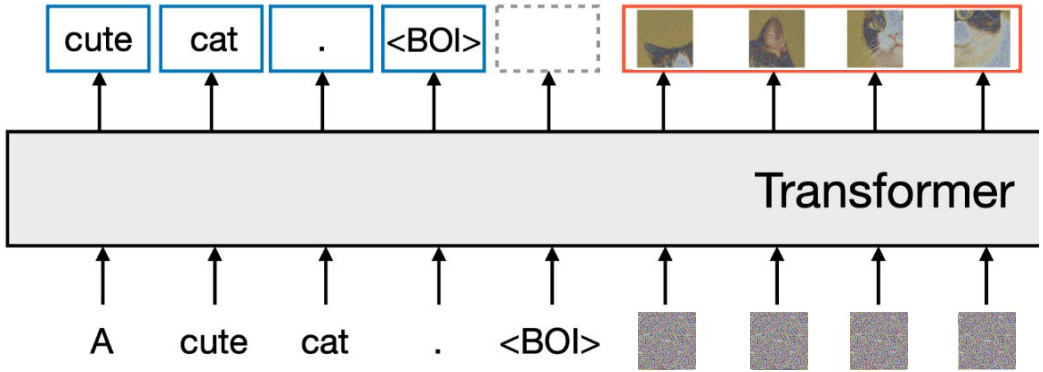
# Transfusion inference



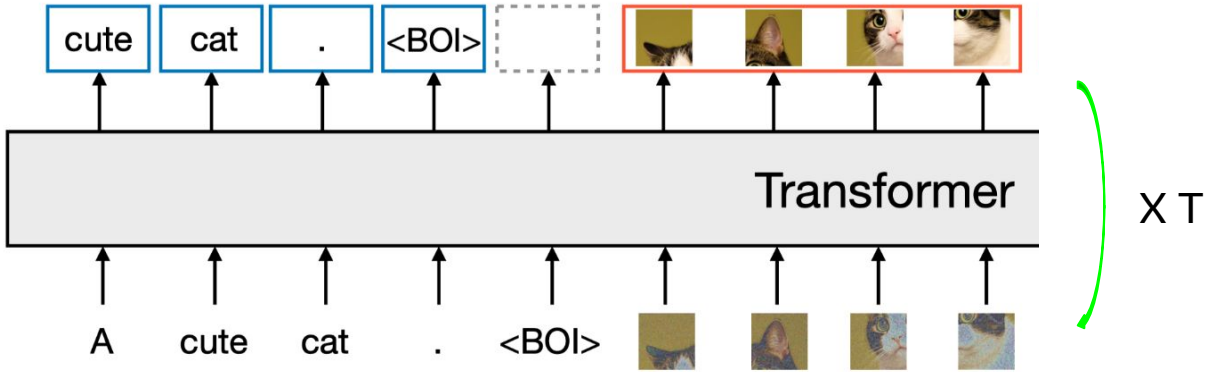
# Transfusion inference



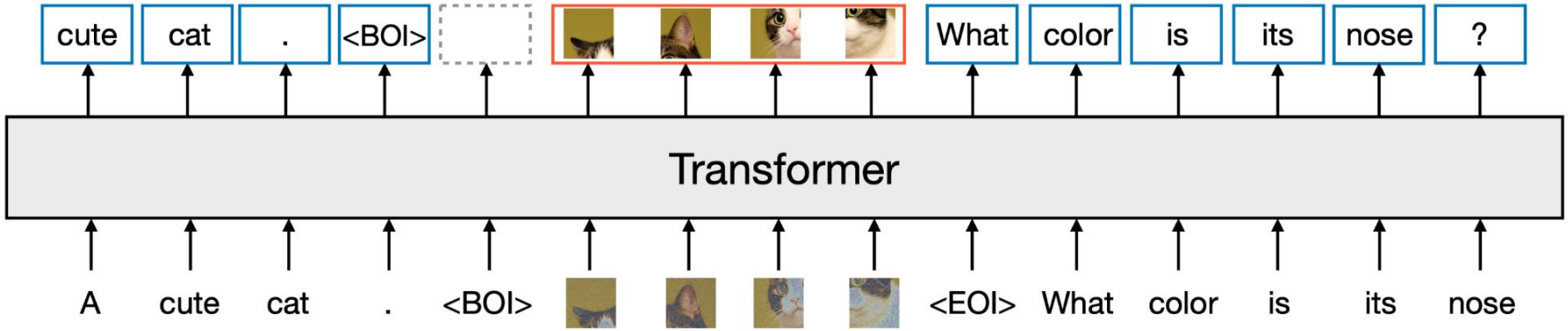
# Transfusion inference



# Transfusion inference



# Transfusion inference




Transfusion can generates arbitrary image-text sequences, enjoy other LM properties, such as KV caching, and in-context learning

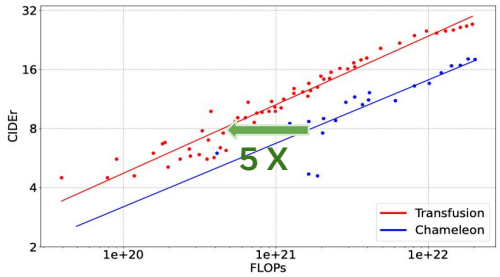
How does transfusion perform on text and image generation?

# Experiment setups

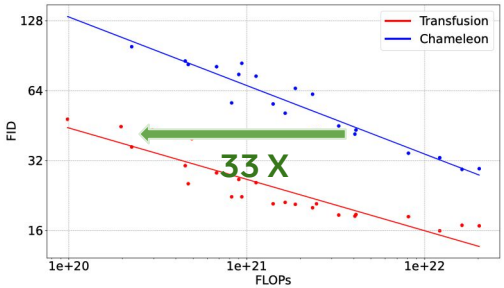
We pretrain transfusion on mixed-modality data, and compare with Chameleon (quantized image tokens) with **controlled tokenizer, data, and model parameters**.

Many model sizes:
163m, 373m, 760m, 1.4b, 7.1b

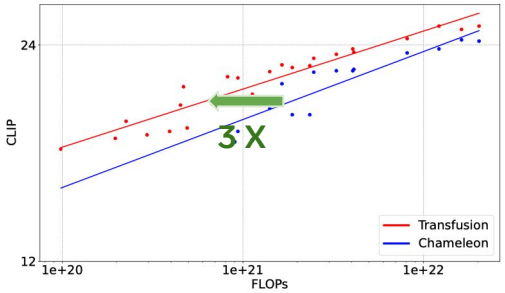
# Result 2: Transfusion vs Chameleon




MS-COCO 5k CIDEr

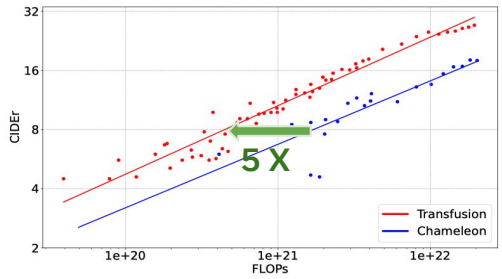



MS-COCO 30k FID

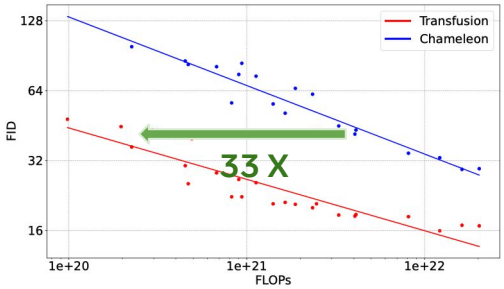



MS-COCO 30k CLIP

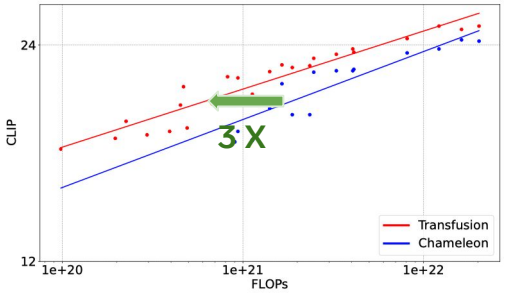
# Result 2: Transfusion vs Chameleon



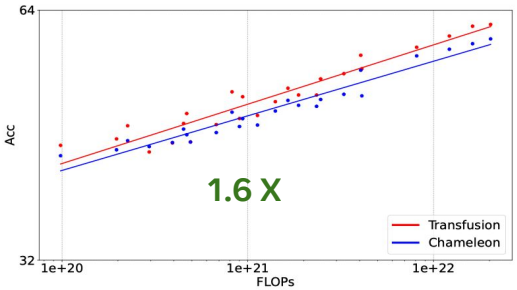

MS-COCO 5k CIDEr



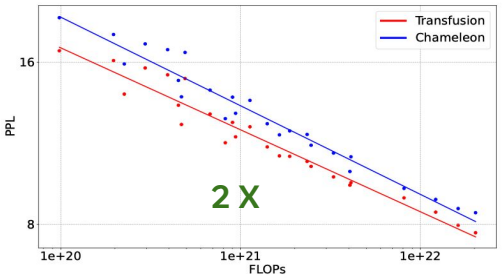

MS-COCO 30k FID



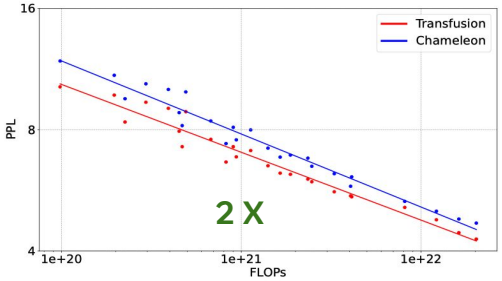

MS-COCO 30k CLIP




Llama 2 Eval Suite Accuracy




C4 Perplexity




Wikipedia Perplexity

**Transfusion scales significantly better than Chameleon in every combination of modalities.**

# Result 2: scale up model

| Model | Model Params | Text Tokens | Images |
|---|---|---|---|
| Transfusion (Ours) | 7.3B | 1.0T | 3.5B |

# Result 2: scale up model

| Model | Model Params | Text Tokens | Images | Llama Acc (↑) |
|---|---|---|---|---|
| Llama 1 [Touvron et al., 2023a] | 7.1B | 1.4T | — | 66.1 |
| Llama 2 [Touvron et al., 2023b] | 7.1B | 2.0T | — | 66.3 |
| Chameleon [Chameleon Team, 2024] | 7.1B | 6.0T | 5.0B | 67.1 |
| Transfusion (Ours) | 7.3B | 1.0T | 3.5B | 66.1 |

# Result 2: scale up model

| Model | Model Params | Text Tokens | Images | Llama Acc (↑) | COCO FID (↓) | Gen Eval (↑) |
|---|---|---|---|---|---|---|
| Llama 1 [Touvron et al., 2023a] | 7.1B | 1.4T | — | 66.1 | — | — |
| Llama 2 [Touvron et al., 2023b] | 7.1B | 2.0T | — | 66.3 | — | — |
| Chameleon [Chameleon Team, 2024] | 7.1B | 6.0T | 5.0B | 67.1 | 26.74 | 0.39 |
| Imagen [Saharia et al., 2022] | 2.6B + 4.7B* | — | 5.0B | — | 7.27 | — |
| Parti [Yu et al., 2022] | 20B | — | 4.8B | — | [r]7.23 | — |
| SD 1.5 [Rombach et al., 2022b] | 0.9B + 0.1B* | — | 4.0B | — | — | 0.43 |
| SD 2.1 [Rombach et al., 2022b] | 0.9B + 0.1B* | — | 2.3B | — | — | 0.50 |
| DALL-E 2 [Ramesh et al., 2022] | 4.2B + 1B* | — | 2.6B | — | 10.39 | 0.52 |
| SDXL [Podell et al., 2023] | 2.6B + 0.8B* | — | 1.6B | — | — | 0.55 |
| DeepFloyd [Stability AI, 2024] | 5.5B + 4.7B* | — | 7.5B | — | 6.66 | 0.61 |
| SD 3 [Esser et al., 2024b] | 8B + 4.7B* | — | [s]2.0B | — | — | 0.68 |
| Transfusion (Ours) | 7.3B | 1.0T | 3.5B | 66.1 | 6.78 | 0.63 |

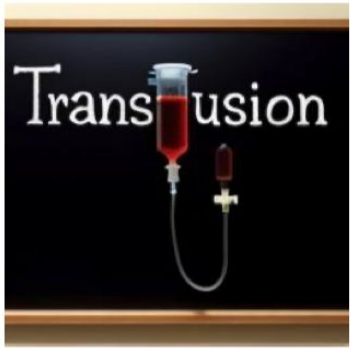
An armchair in the shape of an avocado

A bread, an apple, and a knife on a table

A corgi.

human life depicted entirely out of fractals

A blue jay standing on a large basket of rainbow macarons.




“Transfusion" is written on the blackboard.

A close up photo of a human hand, hand model. High quality

A cloud in the shape of two bunnies playing with a ball. The ball is made of clouds too.


the word ‘START’ on a blue t-shirt

A Dutch still life of an arrangement of tulips in a fluted vase. The lighting is subtle, casting gentle highlights on the flowers and emphasizing their delicate details and natural beauty.

A wall in a royal castle. There are two paintings on the wall. The one on the left a detailed oil painting of the royal raccoon king. The one on the right a detailed oil painting of the royal raccoon queen.

Three spheres made of glass falling into ocean. Water is splashing. Sun is setting.

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

Does transfusion generalize to new modality combinations?

# Generalization text case: image editing

Pretraining:
Text -> text
Text -> image
Image -> text

Finetuning:
Image + text -> image ❓

# Generalization text case: image editing

Pretraining:
Text -> text
Text -> image
Image -> text

Finetuning:
Image + text -> image ✓

**8K examples.**
**Seamlessly edit images.**

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.

# Conclusion

# Conclusion

- How to train a single model to generate SOTA text and image?
- How does transfusion perform on text and image generation?
- Does transfusion generalize to new modality combinations?

# Conclusion

- How to train a single model to generate SOTA text and image?
- ➢ Transfusion: predict the next token and diffuse images in one model.

- How does transfusion perform on text and image generation?

- Does transfusion generalize to new modality combinations?

# Conclusion

- How to train a single model to generate SOTA text and image?

➢ Transfusion: predict the next token and diffuse images in one model.

- How does transfusion perform on text and image generation?

➢ Scales much better than quantized image tokens.

➢ Generates high-quality images, competitive with similar-scale diffusion models, while maintaining zero regret on text generation capabilities.

- Does transfusion generalize to new modality combinations?

# Conclusion

- How to train a single model to generate SOTA text and image?

➢ Transfusion: predict the next token and diffuse images in one model.

- How does transfusion perform on text and image generation?

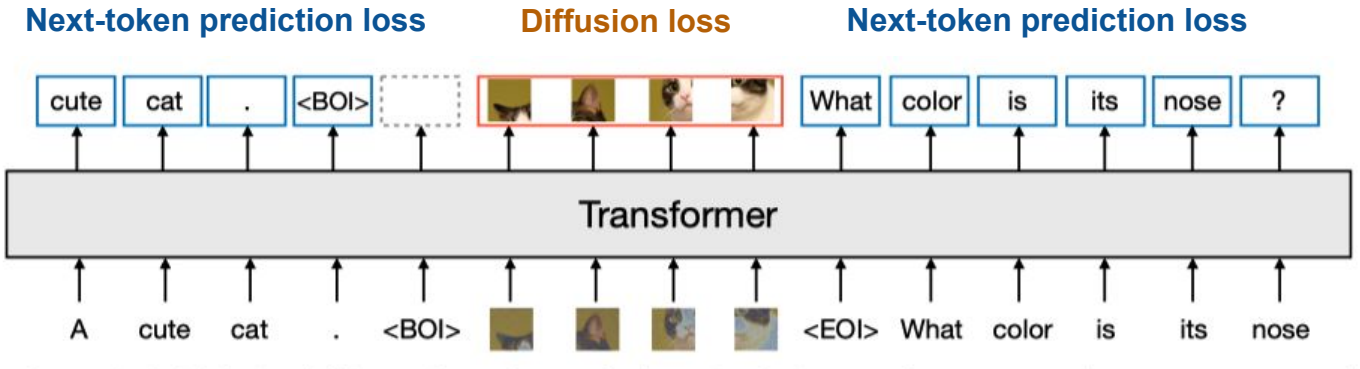
➢ Scales much better than quantized image tokens.

➢ Generates high-quality images, competitive with similar-scale diffusion models, while maintaining zero regret on text generation capabilities.

- Does transfusion generalize to new modality combinations?

➢ Transfusion seamlessly works with any combination of discrete and continuous modalities, such as image editing.

# Questions?




@liliyu_lili

# Transfusion vs Chameleon on text performance

**Deep dive into text performance**

| Model | | Batch | C4 PPL (↓) | Wiki PPL (↓) | Llama Acc (↑) |
|---|---|---|---|---|---|
| Llama 2 | | 1M Text Tokens | 10.1 | 5.8 | 53.7 |
| Transfusion | + Diffusion | + 1M Image Patches | (+0.3) 10.4 | (+0.2) 6.0 | (-1.0) 52.7 |
| Chameleon | + Stability Modifications | 1M Text Tokens | (+0.9) 11.0 | (+0.5) 6.3 | (-1.8) 51.9 |
| | + LM Loss on Image Tokens | + 1M Image Tokens | (+0.8) 11.8 | (+0.5) 6.8 | (-3.0) 48.9 |

Table 4: Performance of the 0.76B Transfusion and Chameleon models on text-only benchmarks, compared to the original Llama 2 recipe.

# Modal Arch: modality-specific encoding+decoding

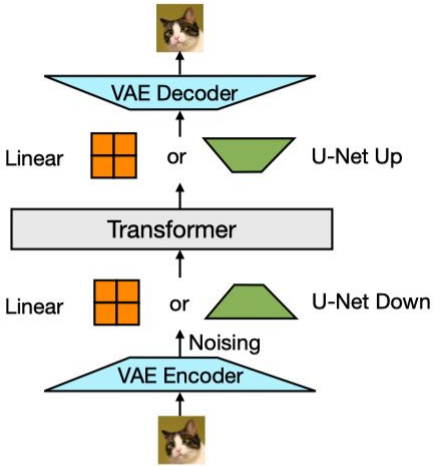



Figure 3: We convert images to and from latent representations using a pretrained VAE, and then into patch representations with either a simple linear layer or U-Net down blocks.

| Model Params | Enc/Dec | Δ Enc/Dec Params | C4 PPL (↓) | Wiki PPL (↓) | Llama Acc (↑) | MS-COCO CDr (↑) | MS-COCO FID (↓) | MS-COCO CLIP (↑) |
|---|---|---|---|---|---|---|---|---|
| 0.16B | Linear | 0.5% | 14.8 | 8.8 | 44.2 | 6.2 | 37.6 | 20.0 |
| | U-Net | 106.1% | 14.4 | 8.5 | 45.7 | 15.3 | 18.8 | 23.9 |
| 0.37B | Linear | 0.4% | 12.0 | 7.0 | 47.9 | 11.1 | 21.5 | 22.4 |
| | U-Net | 71.3% | 11.8 | 6.9 | 48.8 | 21.1 | 18.1 | 24.9 |
| 0.76B | Linear | 0.4% | 10.4 | 6.0 | 51.7 | 16.0 | 20.3 | 24.0 |
| | U-Net | 35.5% | 10.3 | 5.9 | 51.9 | 25.4 | 16.7 | 25.4 |
| 1.4B | Linear | 0.4% | 9.5 | 5.4 | 53.8 | 19.1 | 19.4 | 24.3 |
| | U-Net | 19.3% | 9.4 | 5.4 | 53.4 | 28.1 | 16.6 | 25.7 |
| 7B | Linear | 0.3% | 7.7 | 4.3 | 61.5 | 27.2 | 18.6 | 25.9 |
| | U-Net | 3.8% | 7.8 | 4.3 | 61.1 | 33.7 | 16.0 | 26.5 |

U-net adds fixed parameters, becoming insignificant when transformer grows

Significant improvement on image generation and understanding

# Modal Arch: compress tokens on the fly

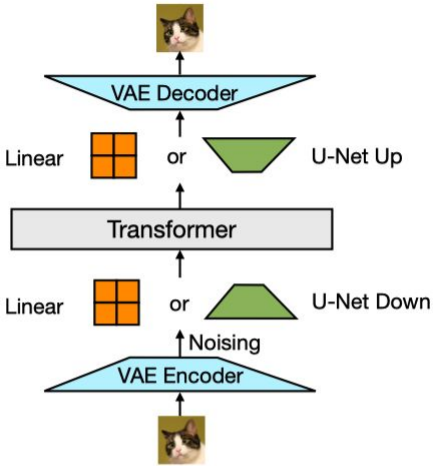



Figure 3: We convert images to and from latent representations using a pretrained VAE, and then into patch representations with either a simple linear layer or U-Net down blocks.

| Enc/Dec | Latent/ Patch | Pixel/ Patch | Patch/ Image |
|---|---|---|---|
| None | 1×1 | 8×8 | 1024 |
| Linear | 2×2 | 16×16 | 256 |
| | 4×4 | 32×32 | 64 |
| | 8×8 | 64×64 | 16 |
| U-Net | 2×2 | 16×16 | 256 |
| | 4×4 | 32×32 | 64 |
| | 8×8 | 64×64 | 16 |

**Each image is presented as much fewer patches with bigger patch size**

# Modal Arch: compress tokens on the fly

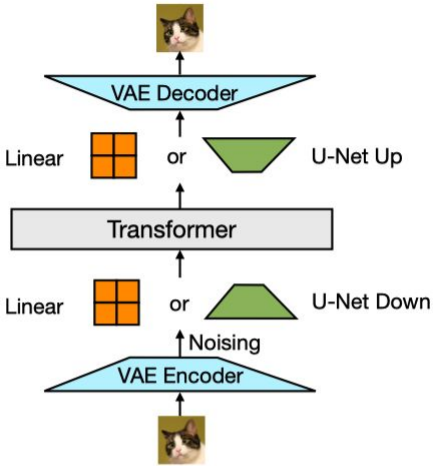
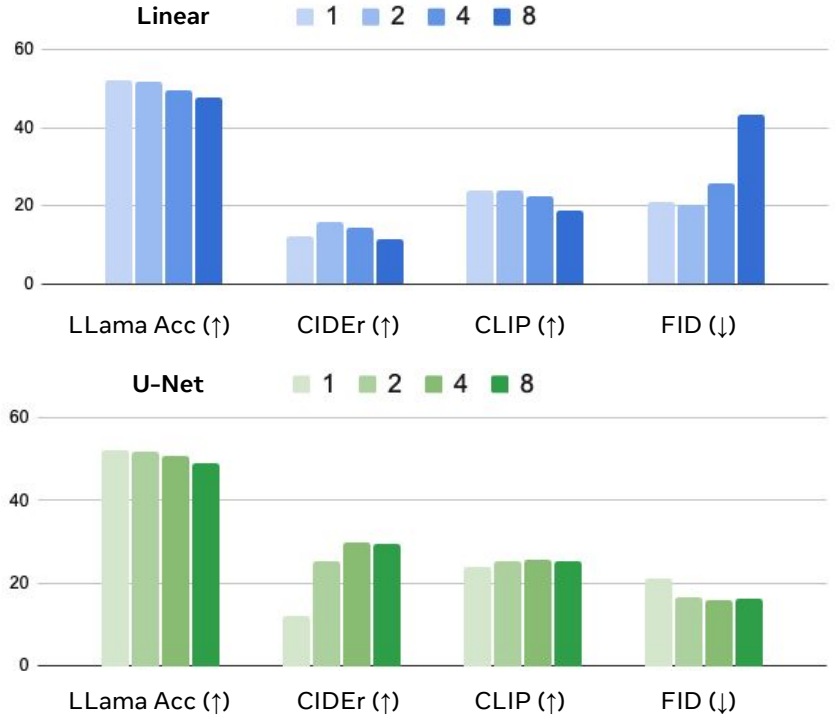



Figure 3: We convert images to and from latent representations using a pretrained VAE, and then into patch representations with either a simple linear layer or U-Net down blocks.




**Transfusion models with U-net can compress each image to just 16 patches.**

# Modal Arch: customized attention mask

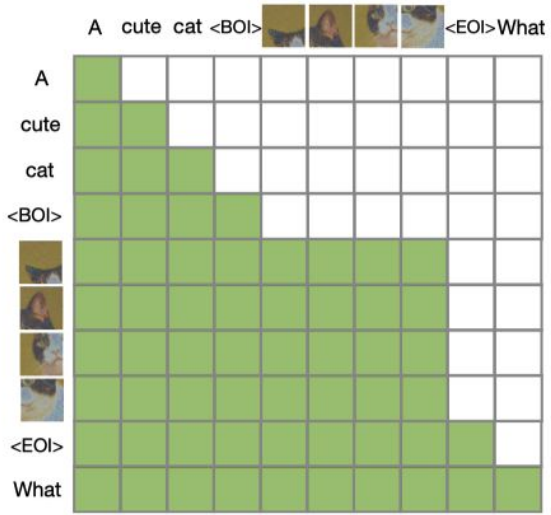
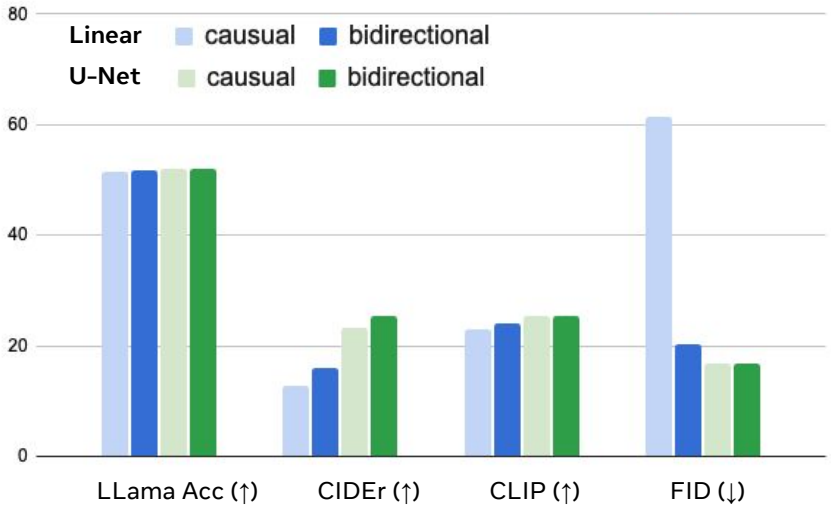



Figure 4: Expanding on the causal mask, Transfusion allows patches of the same image to condition on each other.




**Customized attention mask provide big improvement, particularly for Linear encoder/decoder**