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

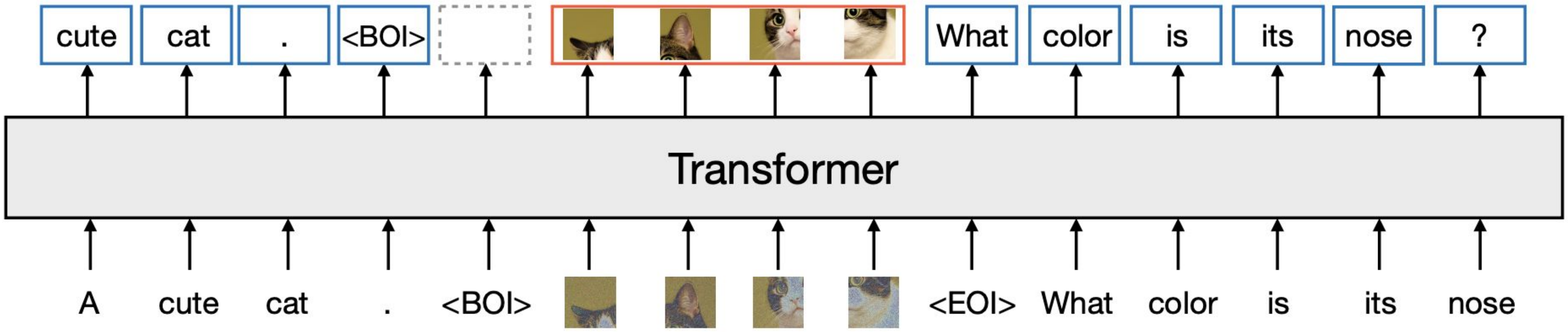
Vision Language Seminar - Sept. 16 2024

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

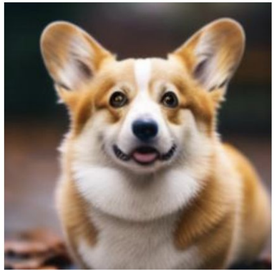
Unifying Text and Image Generation



And it works well!

An armchair in the shape of an avocado

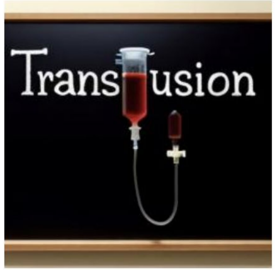
A bread, an apple, and a knife on a table



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.

Some more examples

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

Background: Language Model Loss & Diffusion Loss

Next Token Prediction

$$\mathcal{L}_{\text{LM}} = \mathbb{E}_{y_i}\left[-\log P_\theta(y_i|y_{<i})\right]$$

Diffusion

$$\sqrt{\bar{\alpha}_t} \approx \cos(\tfrac{t}{T} \cdot \tfrac{\pi}{2})$$

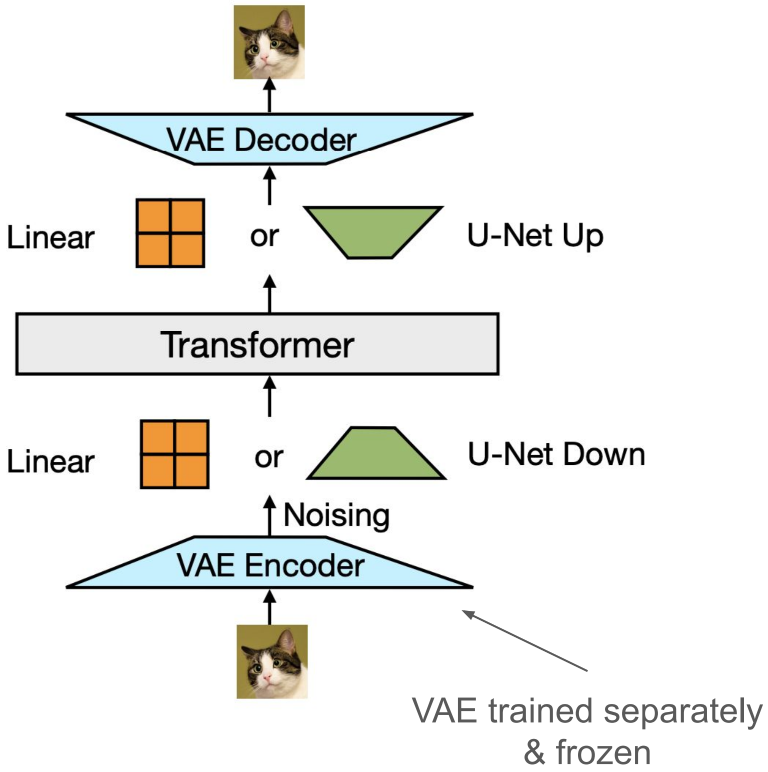
$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$$

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c)||^2\right]$$

Transfusion

Single transformer. Use modality specific components w/ unshared parameters to convert into transformer hidden dimension.

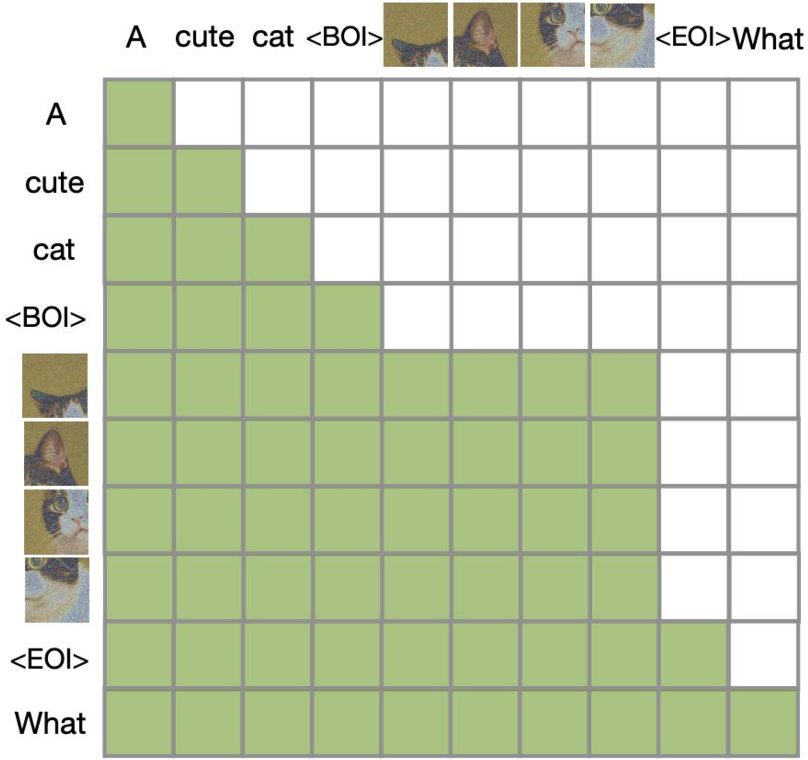
- **Text:** embedding matrix for discrete tokens
- **Images:** VAE transforms images to latents, linear layer/U-Net patchifies and turns latents into embeddings



Transfusion

Attention

- **Text:** causal attention
- **Image:** bidirectional attention, images are not sequential



Transfusion

Training Objective

- For images, add noise ϵ to each input latent x_0 according to the diffusion process to produce x_t
- Apply different losses to text token predictions and image patch predictions
- Use a balancing coefficient and combine losses

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

λ is set to 5 in the paper

Transfusion

Inference

- **LM Mode:** standard sampling, token by token from predicted distribution over vocabulary
- **Diffusion mode:** on <BOI> token, standard diffusion model decoding
  - Append pure noise x_T in the form of n image patches to input sequence and denoise over T steps
  - Given the transformer prediction of noise, denoise x_t to get x_{t-1}
  - Once done, append an <EOI> token to the predicted image and switch back to LM mode

Setup

Data	Sample 0.5T tokens at a 1:1 image-text ratio. 2T text tokens from a diverse distribution of domains 380M images and captions 80% with caption before image, 20% after
VAE	86M parameter w/ CNN encoder and decoder Train for 1M steps $\mathcal{L}_{\text{VAE}} = \mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}} + 0.5\mathcal{L}_{\text{GAN}} + 0.2\mathcal{L}_{\text{ID}} + 0.000001\mathcal{L}_{\text{KL}}$
Model	0.16B, 0.37B, 0.76B, 1.4B, and 7B params to test scaling Greedy decoding for text 1000 diffusion steps for training, 250 steps for inference

Evaluation

Input	Output	Benchmark	Metric
Text	Text	Wikipedia	Perplexity (↓)
		C4	Perplexity (↓)
		Llama 2 Eval Suite	Accuracy (↑)
Image	Text	MS-COCO 5k	CIDEr (↑)
Text	Image	MS-COCO 30k	FID (↓), CLIP (↑)
		GenEval	GenEval score (↑)

Text-only Benchmarks

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	(-2.0) 51.7
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

Training on quantized image tokens degrades text performance more than diffusion on all three benchmarks.
Could be:

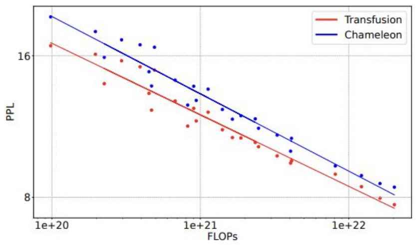
- competition between text and image tokens in the output distribution
- diffusion is more efficient at image generation and requires fewer parameters

All Benchmarks

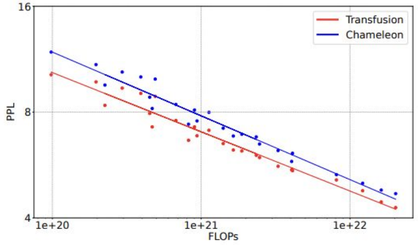
Model	C4 PPL (↓)	Wiki PPL (↓)	Llama Acc (↑)	MS-COCO CDr (↑)	MS-COCO FID (↓)	MS-COCO CLIP (↑)
Transfusion	**7.72**	**4.28**	**61.5**	**27.2**	**16.8**	**25.5**
Chameleon	8.41	4.69	59.1	18.0	29.6	24.3
Parity FLOP Ratio	0.489	0.526	0.600	0.218	0.029	0.319

FLOPs = 6ND
N is num param
D is num tokens processed

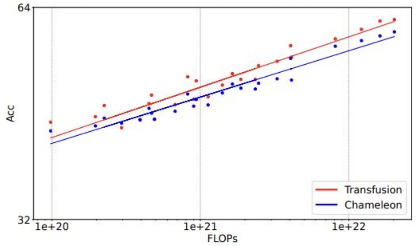
Parity FLOP Ratio: relative Transfusion FLOPs needed to match Chameleon 7B.



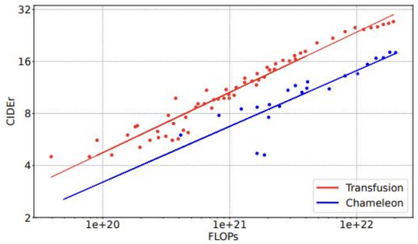
C4 Perplexity



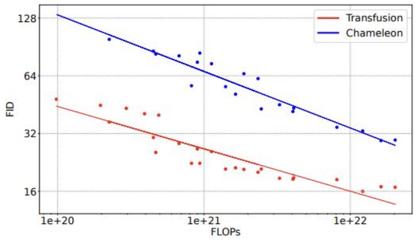
Wikipedia Perplexity



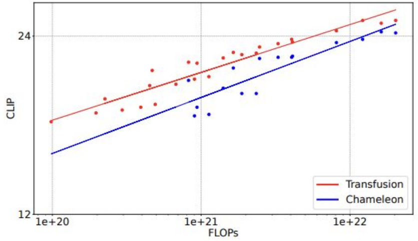
Llama 2 Eval Suite Accuracy



MS-COCO 5k CIDEr



MS-COCO 30k FID



MS-COCO 30k CLIP

Ablations

Attention Masking

Bi-directional attention vs Causal for image patches provides a significant boost in FID (61.3->20.3).

Patch Sizes

Larger patch sizes allow for more images in each training batch and reduce compute, but come at a performance cost. Authors find a good balance at 2*2.

Patch Encoding/Decoding

The model benefits from the inductive biases of a U-Net architecture compared to a Linear layer (possibly hierarchical feature extraction, spatial preservation, etc.).

Ablations

Image Noising

80% of image-caption pairs with caption first (for image generation)

20% of pairs with image first (for image captioning)

For case 2 (image captioning), reduce noising steps to $t = 500$. Significantly improves CIDEr scores (captioning) while having a small effect otherwise.

Comparison with Language & Diffusion Models

Model	Model Params	Text Tokens	Images	Llama Acc (↑)	COCO FID (↓)	Gen Eval (↑)
Llama 1 [Touvron et al., 2023a]	7B	1.4T	—	66.1	—	—
Llama 2 [Touvron et al., 2023b]	7B	2.0T	—	66.3	—	—
Chameleon [Chameleon Team, 2024]	7B	6.0T	3.5B	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

Image Editing

Fine-tuned with only 8k image editing samples.

input image, edit prompt -> output image

Powerful generalization capabilities!

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 / Comments

- Lots of comments about lack of evaluation on more complex visual understanding and reasoning benchmarks (e.g. TextVQA, VSR, VQAv2)
- **Joint architecture**
  - Annya: What are the shortcomings of this approach of having a single joint model on two objectives? Why aren't all multimodal approaches conducted in this same way?
  - Junyi: How does Transfusion handle the integration and potential interference between the language modeling and diffusion objectives during training, and what strategies could be employed to further optimize their coexistence for even better multimodal performance?

Questions / Comments

- Rudy: The results in Table 9 appear to imply that Transfusion's attempt at being a jack of all trades make it a master of none. One could imagine an alternate universe where the variety of data and tasks would synergize and result in even greater improvements. Why is this not the case, does optimizing a representation for within-modality generation hurt its ability to be useful as a conditioning variable for the other modality?
- Ren: Why does Transfusion outperform Chameleon on text-only tasks? Can the diffusion objective for image tokens really account for this difference?