



# Any Task as Distribution Modeling

Paper Reading by Zhiying Lu

2025.03.18

智能多媒体内容计算实验室  
Intelligent Multimedia Content Computing Lab



- 作者介绍
- 背景介绍
- 本文方法
- 实验效果
- 总结反思

# 作者介绍

3

Zijie Li<sup>1\*</sup> Henry Li<sup>2\*</sup> Yichun Shi<sup>3</sup> Amir Barati Farimani<sup>1</sup> Yuval Kluger<sup>2</sup>  
 Linjie Yang<sup>3</sup> Peng Wang<sup>3</sup>

<sup>1</sup>Carnegie Mellon University <sup>2</sup>Yale University <sup>3</sup>Bytedance Seed

[Project website](#)



Peng Wang

[University of California, Los Angeles / Bytedance, Inc](#)  
在 ucla.edu 的电子邮件经过验证 - 首页

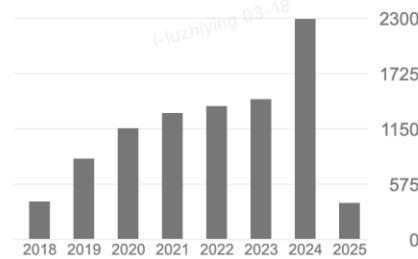
Computer Vision Machine Learning Image Processing

关注

引用次数

查看全部

	总计	2020 年至今
引用	9772	7983
h 指数	42	38
i10 指数	80	74



标题	引用次数	年份
----	------	----

[Seedream 2.0: A Native Chinese-English Bilingual Image Generation Foundation Model](#)

L Gong, X Hou, F Li, L Li, X Lian, F Liu, L Liu, W Liu, W Lu, Y Shi, S Sun, ...  
arXiv preprint arXiv:2503.07703

引用次数  
年份

2025

[Box detection for object attachment](#)

S Cheng, P Wang  
US Patent 12,190,554

2025

[MBA-SLAM: Motion Blur Aware Dense Visual SLAM with Radiance Fields Representation](#)

P Wang, L Zhao, Y Zhang, S Zhao, P Liu  
arXiv preprint arXiv:2411.08279

2024

[SeedEdit: Align Image Re-Generation to Image Editing](#)

Y Shi, P Wang, W Huang  
arXiv preprint arXiv:2411.06686

2 2024

[Visual text generation in the wild](#)

Y Zhu, J Liu, F Gao, W Liu, X Wang, P Wang, F Huang, C Yao, Z Yang  
European Conference on Computer Vision, 89-106

5 2024

[Pixel perspective estimation and refinement in an image](#)

P Wang  
US Patent App. 18/175,738

2024

[Image rendering method, apparatus, electronic device and storage medium](#)

P Wang, G Wang, Z Wang  
US Patent App. 18/434,499

2024

Proceedings of the National Academy of Sciences 118 (22), e2100293118

[Variational diffusion autoencoders with random walk sampling](#)

H Li, O Lindenbaum, X Cheng, A Cloninger  
Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23 ...

16 2020



计算实验室

开放获取的出版物数量 [查看全部](#)

1 篇文章 [28 篇文章](#)

无法查看的文章 [可查看的文章](#)

根据资助方的强制性开放获取政策

合著者 [查看全部](#)

Ruigang Yang  
Baidu Research/Professor of Co...

>

...tent Computing Lab

# 作者介绍

4

## Large Language Diffusion Models

Shen Nie<sup>1\*†</sup> Fengqi Zhu<sup>1\*†</sup> Zebin You<sup>1†</sup> Xiaolu Zhang<sup>2‡</sup> Jingyang Ou<sup>1</sup> Jun Hu<sup>2‡</sup> Jun Zhou<sup>2</sup>  
Yankai Lin<sup>1†</sup> Ji-Rong Wen<sup>1</sup> Chongxuan Li<sup>1‡¶</sup>



Shen Nie

[Renmin University of China](#)  
Verified email at ruc.edu.cn  
generative model

FOLLOW

Cited by

All  
Citations  
536

Since 2020  
536

Chongxuan Li

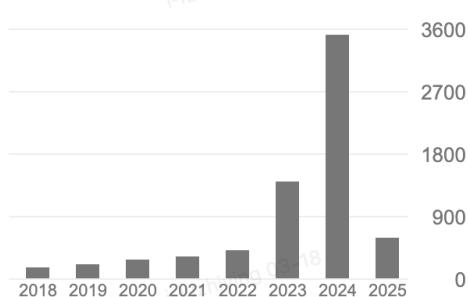
关注

Associate Professor (tenure track), [Renmin University of China](#)  
在 ruc.edu.cn 的电子邮件经过验证 - 首页

Machine Learning Generative Models Deep Learning

引用次数

总计  
引用  
7027  
h 指数  
34  
i10 指数  
51  
2020 年至今  
6560  
33  
51



标题

引用次数 年份

Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps

C Lu, Y Zhou, F Bao, J Chen, C Li, J Zhu  
NeurIPS 2022

1298 2022

Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation

Z Wang, C Lu, Y Wang, F Bao, C Li, H Su, J Zhu  
Advances in Neural Information Processing Systems 36, 8406-8441

805 2023

Towards better analysis of deep convolutional neural networks

M Liu, J Shi, Z Li, C Li, J Zhu, S Liu  
IEEE transactions on visualization and computer graphics 23 (1), 91-100

627 2016

Triple generative adversarial nets

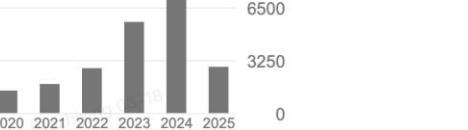
C Li, T Xu, J Zhu, B Zhang  
Advances in neural information processing systems 30  
arXiv preprint arXiv:2303.18223

573 2017

开放获取的出版物数量 [查看全部](#)

2 篇文章 [34 篇文章](#)

无法查看的文  
章 可查看的文  
章



A survey on large language model based autonomous agents

L Wang, C Ma, X Feng, Z Zhang, H Yang, J Zhang, Z Chen, J Tang, ...  
Frontiers of Computer Science 18 (6), 186345

1082 2024

Vips: a vision-based page segmentation algorithm

D Cai, S Yu, JR Wen, WY Ma

997 2003

实验室



- 作者介绍
- 背景介绍
- 本文方法
- 实验效果
- 总结反思



# 背景知识

6

- 扩散模型
  - 图像Diffusion, Flow Matching
  - Diffusion Transformer (DiT), MM-DiT
  - Text Diffusion
- 生成理解统一模型
  - 自回归式
  - tokenizer



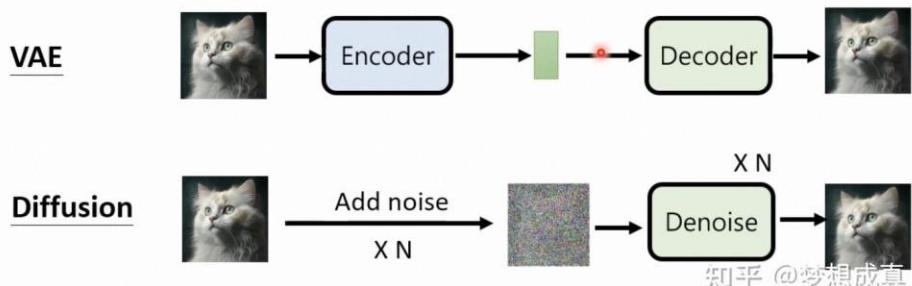
# 模型设计角度

7

- 所有模型都可以视为：Encoder->Predictor->Decoder
- **视觉生成模型**：VAE编码器压缩到隐空间，在隐空间结合DiT进行分布变换，利用VAE解码器来解码
- **视觉理解模型**：PatchEmb得到连续空间的视觉表示，利用ViT进行分布变换，利用线性层对应到类别
- **文本模型**：Tokenizer编码到离散空间，利用Transformer进行分布变换，利用Codebook解码得到文字

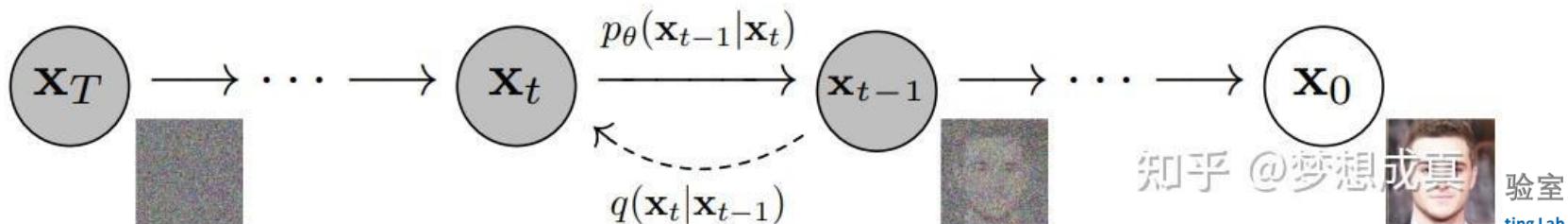
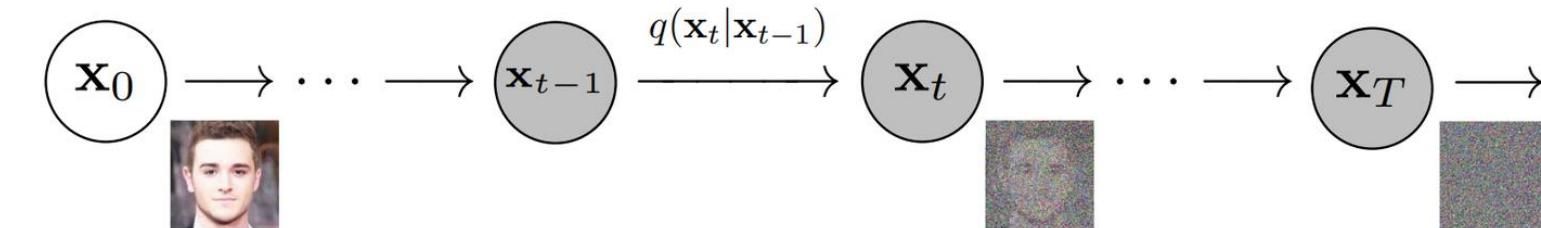
# 扩散模型

8



知乎 @梦想成真

- 生成模型本身做的是一个**分布变换**
- 首先训练一个VAE隐空间，压缩特征空间和通道，得到成对的编码器和解码器
- 在隐空间中加噪训练-去噪推理，扩散模型作为**每一步的噪声预测器**，**预测方向大小**
  - 训练时，将任意图像逐步加噪声直到标准高斯分布
  - 推理时，将标准高斯分布转换成目标图像



知乎 @梦想成真 验室 ting Lab



# 扩散模型v.s.流匹配

9

- DDPM和Flow Matching都是利用**逐步逼近（积分）**进行分布变换
- DDPM预测每一步**高斯噪声**，FM预测每一步的**速度场**
- FM收敛效率更高，稳定性和效果更好，目前为主流

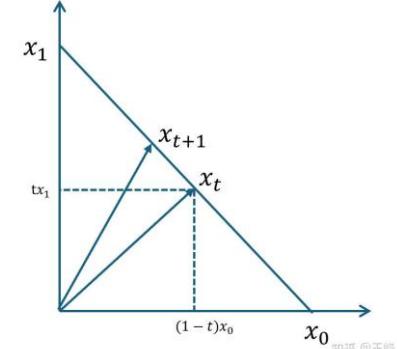
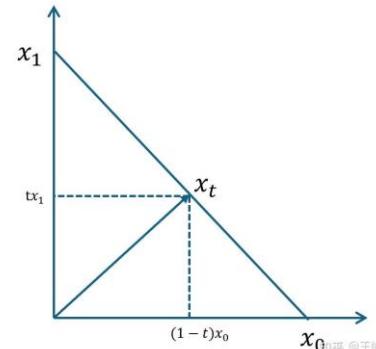
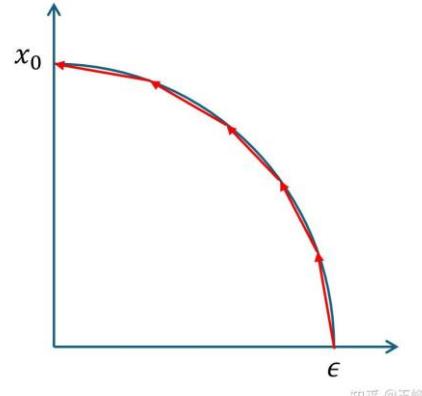
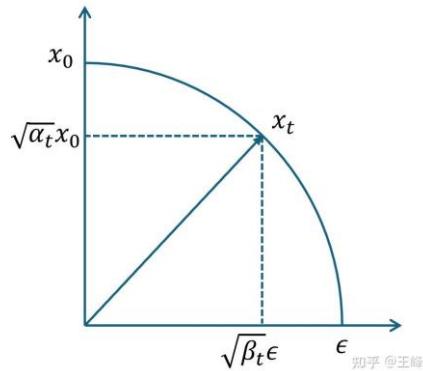
$$\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon,$$

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{\beta_t} \epsilon,$$

$$\beta_t = 1 - \alpha_t,$$

$$p_{\theta}(\mathbf{x}) = \int p_{\theta}(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \quad \dot{\mathbf{x}}_t = \mathbf{v}(\mathbf{x}_t, t),$$
$$\mathbf{v}(\mathbf{x}_t, t) = \dot{\alpha}_t \mathbf{x} + \dot{\sigma}_t \epsilon.$$

$$x_t = tx_1 + (1-t)x_0,$$

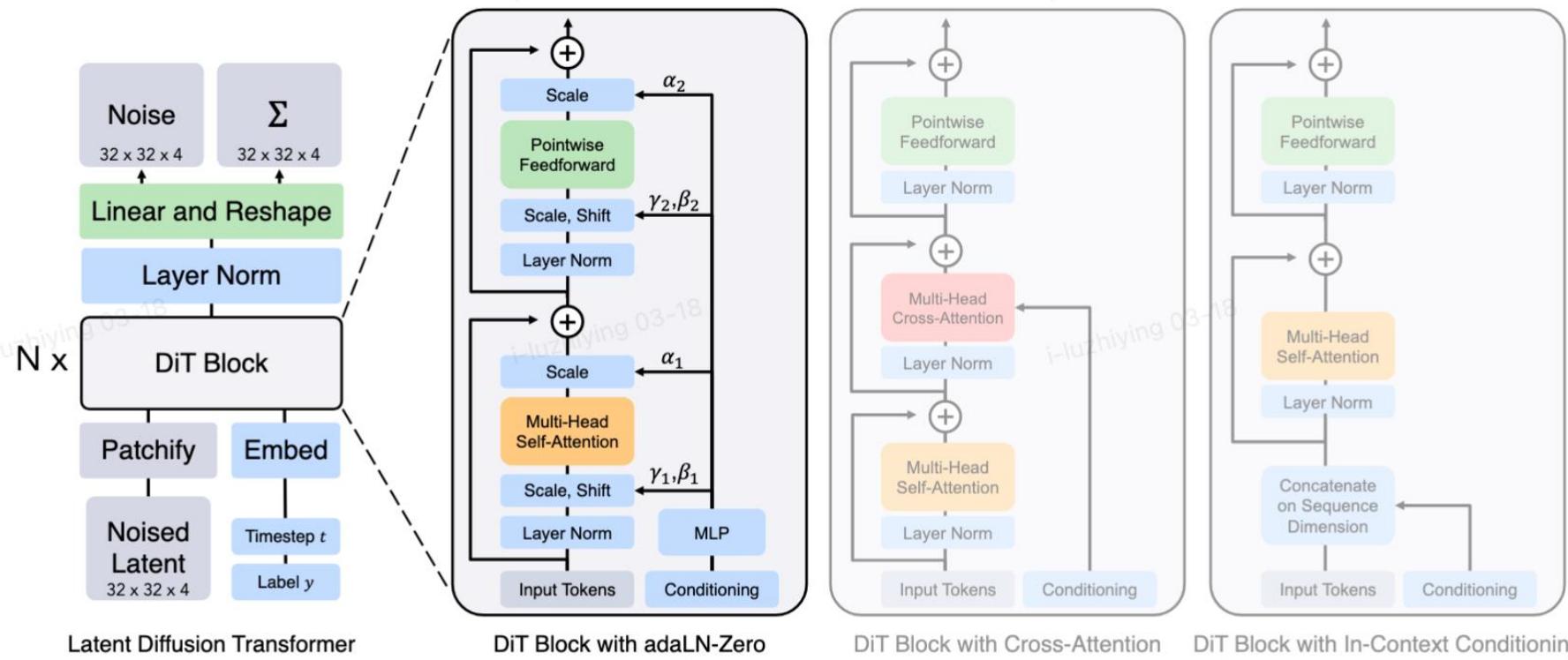




# Diffusion Transformer

10

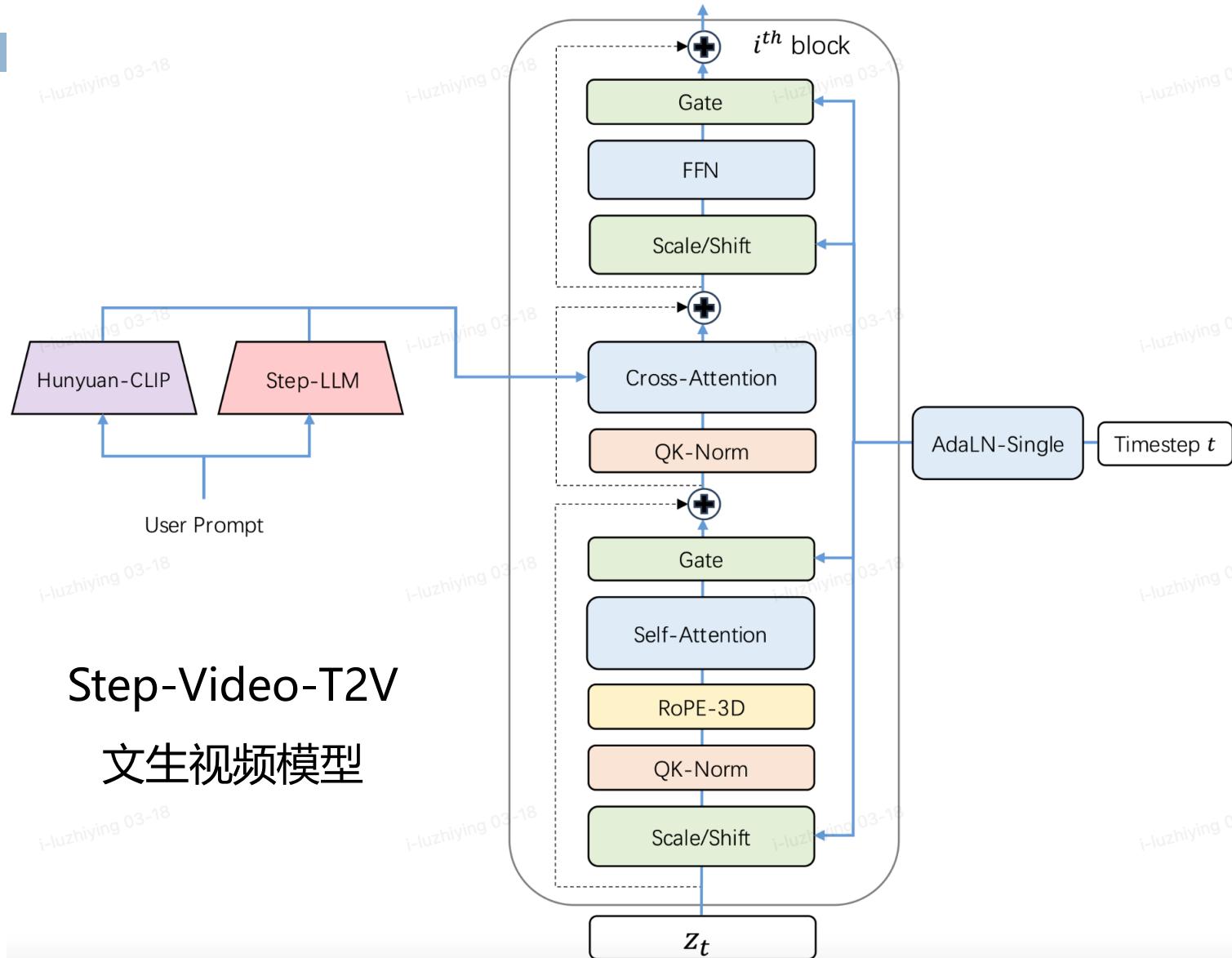
- DiT根据**输入特征**, **时间步** (扩散过程进行到哪一步) , **条件信号**, 进行加噪/去噪的**幅度和方向**的预测 (FM的速度场同理)
- 以各种信号为条件的条件分布变换





# Diffusion Transformer

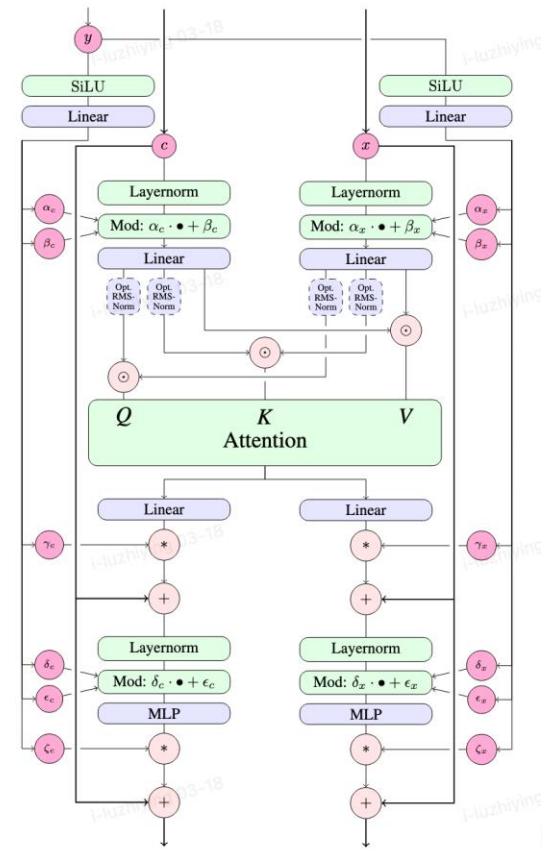
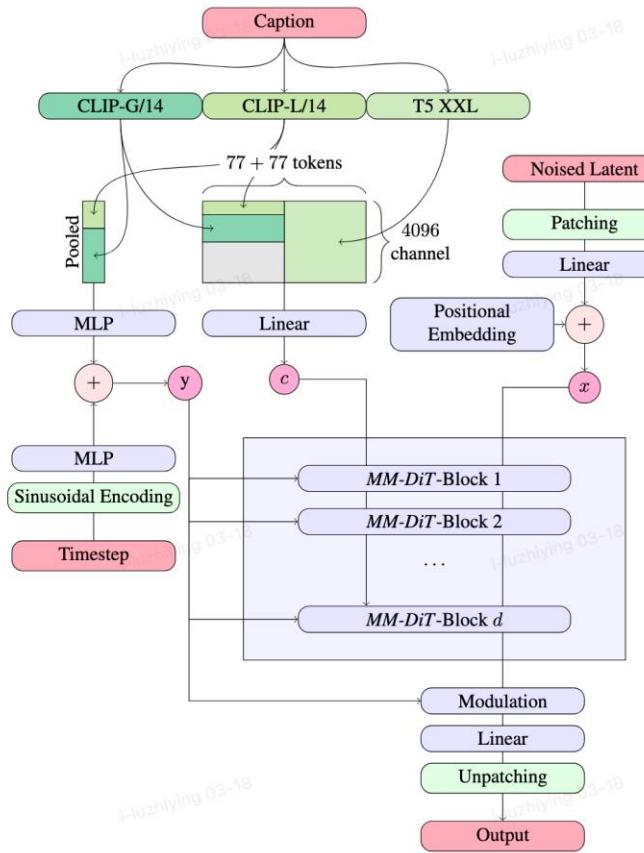
11



# Diffusion Transformer

12

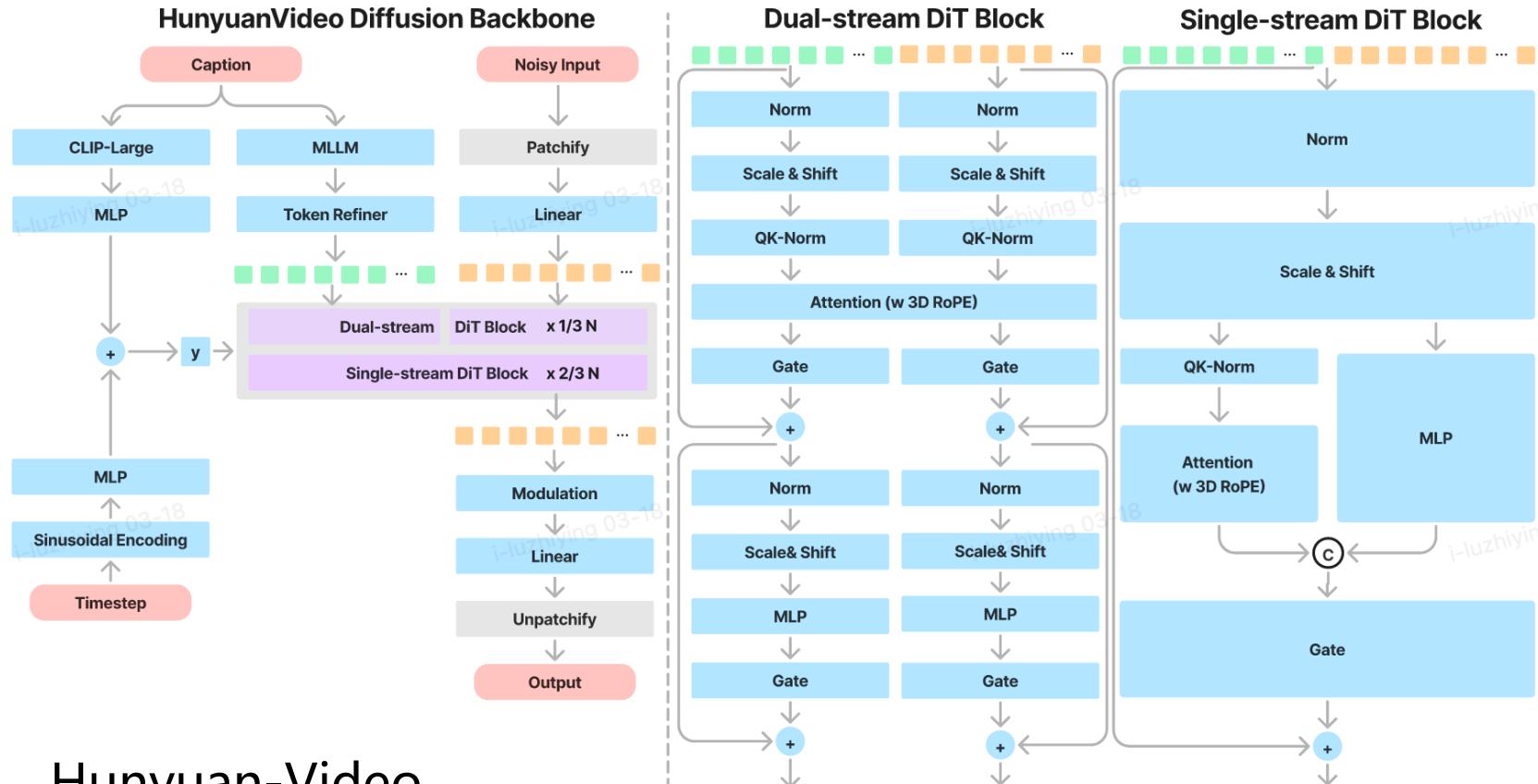
- MM-DiT来自于Stable Diffusion3，将文本和视觉特征concat做自注意力，但具有各自的FFN--联合交互，分开表征





# Diffusion Transformer

13



Hunyuan-Video

文生视频模型



# Text Diffusion

14

- 传统的文本模型依赖于离散的码本 (Vicuna码本3万, CLIP码本5万)
- 文本建模大多采用自回归, 过一次模型只能输出一个token, 慢且单向
- 最新的masked diffusion model允许一次性预测多个token, 快且双向



# Text Diffusion

15

## Large Language Diffusion Models

Shen Nie<sup>1 \* †</sup> Fengqi Zhu<sup>1 \* †</sup> Zebin You<sup>1 †</sup> Xiaolu Zhang<sup>2 ‡</sup> Jingyang Ou<sup>1</sup> Jun Hu<sup>2 ‡</sup> Jun Zhou<sup>2</sup>  
Yankai Lin<sup>1 ‡</sup> Ji-Rong Wen<sup>1</sup> Chongxuan Li<sup>1 ‡ ¶</sup>

$$\underbrace{\max_{\theta} \mathbb{E}_{p_{\text{data}}(x)} \log p_{\theta}(x) \Leftrightarrow \min_{\theta} \text{KL}(p_{\text{data}}(x) || p_{\theta}(x))}_{\text{Generative modeling principles}}.$$

$$p_{\theta}(x) = p_{\theta}(x^1) \underbrace{\prod_{i=2}^L p_{\theta}(x^i | x^1, \dots, x^{i-1})}_{\text{Autoregressive formulation}},$$



# Text Diffusion

16

(a) Mask all tokens independently



Mask predictor



$\boxtimes$  Mask token

$\downarrow$  Remask

Non-mask token

$\downarrow$  Random mask

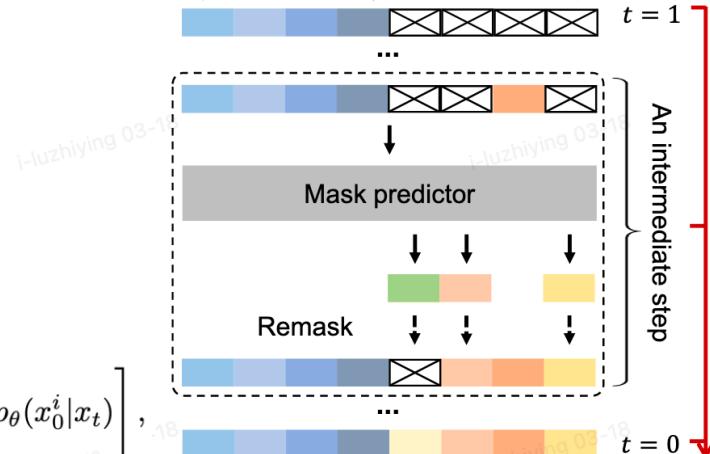
(b) Prompt Response



Mask predictor



(c) Prompt Response



$$\mathcal{L}(\theta) \triangleq -\mathbb{E}_{t,x_0,x_t} \left[ \frac{1}{t} \sum_{i=1}^L \mathbf{1}[x_t^i = M] \log p_\theta(x_0^i | x_t) \right],$$

Figure 2. A Conceptual Overview of LLaDA. (a) Pre-training. LLaDA is trained on text with random masks applied independently to all tokens at the same ratio  $t \sim U[0, 1]$ . (b) SFT. Only response tokens are possibly masked. (c) Sampling. LLaDA simulates a diffusion process from  $t = 1$  (fully masked) to  $t = 0$  (unmasked), predicting all masks simultaneously at each step with flexible remask strategies.

- 训练时间步  $t=0 \sim 1$ , 也是每个token被mask的概率
- loss只在mask token上计算, 一旦被mask就不会再变动
- 从完全无mask训练到完全被mask
- 解码时从response完全mask到完全无mask

$$q_{t|0}(x_t | x_0) = \prod_{i=1}^L q_{t|0}(x_t^i | x_0^i),$$

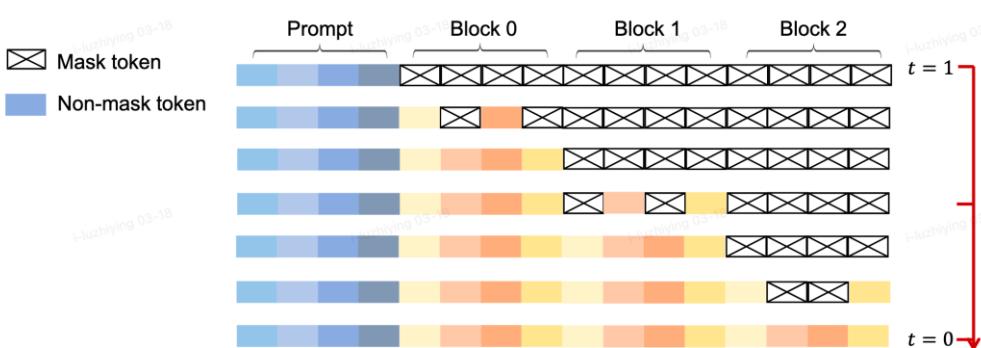
$$q_{t|0}(x_t^i | x_0^i) = \begin{cases} 1-t, & x_t^i = x_0^i, \\ t, & x_t^i = M. \end{cases}$$

$$q_{s|t}(x_s^i | x_t) = \prod_{i=1}^L q_{s|t}(x_s^i | x_t),$$

$$q_{s|t}(x_s^i | x_t) = \begin{cases} 1, & x_t^i \neq M, x_s^i = x_t^i, \\ \frac{s}{t}, & x_t^i = M, x_s^i = M, \\ \frac{t-s}{t} q_{0|t}(x_s^i | x_t), & x_t^i = M, x_s^i \neq M, \\ 0, & \text{otherwise}. \end{cases}$$

# Text Diffusion

17



User  
LLaDA

What are the benefits of regular exercise for physical and mental health?

pling.

Regular exercise can improve physical health by strengthening muscles and bones, improving cardiovascular health, and reducing the risk of chronic diseases such as diabetes and heart disease. It can also improve mental health by reducing stress and anxiety, improving mood, and enhancing cognitive function. Additionally, exercise can improve sleep quality and increase energy levels.

User  
LLaDA

I want to have a pet, but I don't know whether to get a cat or a dog. Do you have any suggestions?

It really depends on your preferences, lifestyle, and the type of pet you want. If you want an independent pet and have a busy schedule, a cat is a good choice. If you prefer a companion and have a more active lifestyle, a dog might be a better fit. Have you considered your options?

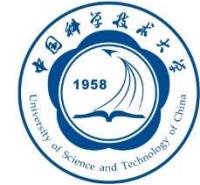
User  
LLaDA

I'm very busy at work, so it seems like a cat would be a more suitable option. Do you have any suggestions for getting a cat?

Sure! First, make sure you have enough space in your home and provide food, water, and toys for your cat. Also, research different breeds and personalities, and choose cats that are suitable for your lifestyle. Lastly, make sure you have the necessary vaccinations and insurance in place before you get a pet.

# Text Diffusion

Table 3. Comparison in the Poem Completion Task.



	Forward	Reversal
GPT-4o (2024-08-06)	<b>82.7</b>	34.3
Qwen2.5 7B Instruct	75.9	38.0
LLaDA 8B Instruct	48.8	<b>42.4</b>

Table 1. Benchmark Results of Pre-trained LLMs. \* indicates that LLaDA 8B Base, LLaMA2 7B Base, and LLaMA3 8B Base are evaluated under the same protocol, detailed in Appendix B.5. Results indicated by <sup>†</sup> and <sup>¶</sup> are sourced from Chu et al. (2024); Yang et al. (2024) and Bi et al. (2024) respectively. The numbers in parentheses represent the number of shots used for evaluation. “-” indicates unknown data.

	LLaDA 8B*	LLaMA3 8B*	LLaMA2 7B*	Qwen2 7B <sup>†</sup>	Qwen2.5 7B <sup>†</sup>	Mistral 7B <sup>†</sup>	Deepseek 7B <sup>¶</sup>
Model	Diffusion	AR	AR	AR	AR	AR	AR
Training tokens							
	2.3T	15T	2T	7T	18T	-	2T
General Tasks							
MMLU	<b>65.9</b> (5)	65.4 (5)	45.9 (5)	70.3 (5)	74.2 (5)	64.2 (5)	48.2 (5)
BBH	49.8 (3)	<b>57.6</b> (3)	37.3 (3)	62.3 (3)	70.4 (3)	56.1 (3)	39.5 (3)
ARC-C	47.9 (0)	<b>53.1</b> (0)	46.3 (0)	60.6 (25)	63.7 (25)	60.0 (25)	48.1 (0)
Hellaswag	72.5 (0)	<b>79.1</b> (0)	76.0 (0)	80.7 (10)	80.2 (10)	83.3 (10)	75.4 (0)
TruthfulQA	<b>46.4</b> (0)	44.0 (0)	39.0 (0)	54.2 (0)	56.4 (0)	42.2 (0)	-
WinoGrande	74.8 (5)	<b>77.3</b> (5)	72.5 (5)	77.0 (5)	75.9 (5)	78.4 (5)	70.5 (0)
PIQA	74.4 (0)	<b>80.6</b> (0)	79.1 (0)	-	-	-	79.2 (0)
Mathematics & Science							
GSM8K	<b>70.7</b> (4)	53.1 (4)	14.3 (4)	80.2 (4)	85.4 (4)	36.2 (4)	17.4 (8)
Math	<b>27.3</b> (4)	15.1 (4)	3.2 (4)	43.5 (4)	49.8 (4)	10.2 (4)	6.0 (4)
GPQA	<b>26.1</b> (5)	25.9 (5)	25.7 (5)	30.8 (5)	36.4 (5)	24.7 (5)	-
Code							
HumanEval	33.5 (0)	<b>34.2</b> (0)	12.8 (0)	51.2 (0)	57.9 (0)	29.3 (0)	26.2 (0)
HumanEval-FIM	<b>73.8</b> (2)	73.3 (2)	26.9 (2)	-	-	-	-
MBPP	38.2 (4)	<b>47.4</b> (4)	18.4 (4)	64.2 (0)	74.9 (0)	51.1 (0)	39.0 (3)
Chinese							
CMMLU	<b>69.9</b> (5)	50.7 (5)	32.5 (5)	83.9 (5)	-	-	47.2 (5)
C-Eval	<b>70.5</b> (5)	51.7 (5)	34.0 (5)	83.2 (5)	-	-	45.0 (5)



# Text Diffusion

19

**Table 2. Benchmark Results of Post-trained LLMs.** LLaDA only employs an SFT procedure while other models have extra reinforcement learning (RL) alignment. \* indicates that LLaDA 8B Instruct, LLaMA2 7B Instruct, and LLaMA3 8B Instruct are evaluated under the same protocol, detailed in Appendix B.5. Results indicated by <sup>†</sup> and <sup>¶</sup> are sourced from Yang et al. (2024) and Bi et al. (2024) respectively. The numbers in parentheses represent the number of shots used for in-context learning. “-” indicates unknown data.

	LLaDA 8B*	LLaMA3 8B*	LLaMA2 7B*	Qwen2 7B <sup>†</sup>	Qwen2.5 7B <sup>†</sup>	Gemma2 9B <sup>†</sup>	Deepseek 7B <sup>¶</sup>
Model	Diffusion	AR	AR	AR	AR	AR	AR
Training tokens	2.3T	15T	2T	7T	18T	8T	2T
Post-training Alignment pairs							
SFT	SFT+RL	SFT+RL	SFT+RL	SFT+RL	SFT+RL	SFT+RL	SFT+RL
4.5M	-	-	0.5M + -	1M + 0.15M	-	-	1.5M + -
General Tasks							
MMLU	65.5 (5)	<b>68.4</b> (5)	44.1 (5)	-	-	-	49.4 (0)
MMLU-pro	37.0 (0)	<b>41.9</b> (0)	4.6 (0)	44.1 (5)	56.3 (5)	52.1 (5)	-
Hellaswag	74.6 (0)	<b>75.5</b> (0)	51.5 (0)	-	-	-	68.5 (-)
ARC-C	<b>88.5</b> (0)	82.4 (0)	57.3 (0)	-	-	-	49.4 (-)
Mathematics & Science							
GSM8K	<b>78.6</b> (4)	78.3 (4)	29.0 (4)	85.7 (0)	91.6 (0)	76.7 (0)	63.0 (0)
Math	26.6 (0)	<b>29.6</b> (0)	3.8 (0)	52.9 (0)	75.5 (0)	44.3 (0)	15.8 (0)
GPQA	31.8 (5)	<b>31.9</b> (5)	28.4 (5)	34.3 (0)	36.4 (0)	32.8 (0)	-
Code							
HumanEval	47.6 (0)	<b>59.8</b> (0)	16.5 (0)	79.9 (0)	84.8 (0)	68.9 (0)	48.2 (-)
MBPP	34.2 (4)	<b>57.6</b> (4)	20.6 (4)	67.2 (0)	79.2 (0)	74.9 (0)	35.2 (-)



# 生成理解统一模型

20

- 大部分基于纯自回归或者自回归语言 + 扩散视觉解码

**Multimodal Understanding**  
(Captioning, VQA ...)



Q1: Please describe this image in detail.  
Q2: Is there a rainbow in this image?

**Visual Generation**  
(Text-to-Image Generation / Text-guided Inpainting and Extrapolation)

A punk rock frog in a studded leather jacket shouting into a microphone while standing on a boulder.



a dog sitting on the bench.



a vibrant hot air balloon floats over a clear lake.

**Mixed-modality generation**  
(Video keyframe generation with text descriptions)



Slicing avocado.

Text Tokenizer & Image Tokenizer



**Show-O (Causal & Full Attention)**



Text De-Tokenizer & Image De-Tokenizer

A1: The image features a young girl sitting on the grass, surrounded by a colourful backdrop. She is holding a ...



A2: Yes, there is a rainbow in the image, as the girl is painting a rainbow on the canvas.

a woman is cutting an avocado with a knife ...

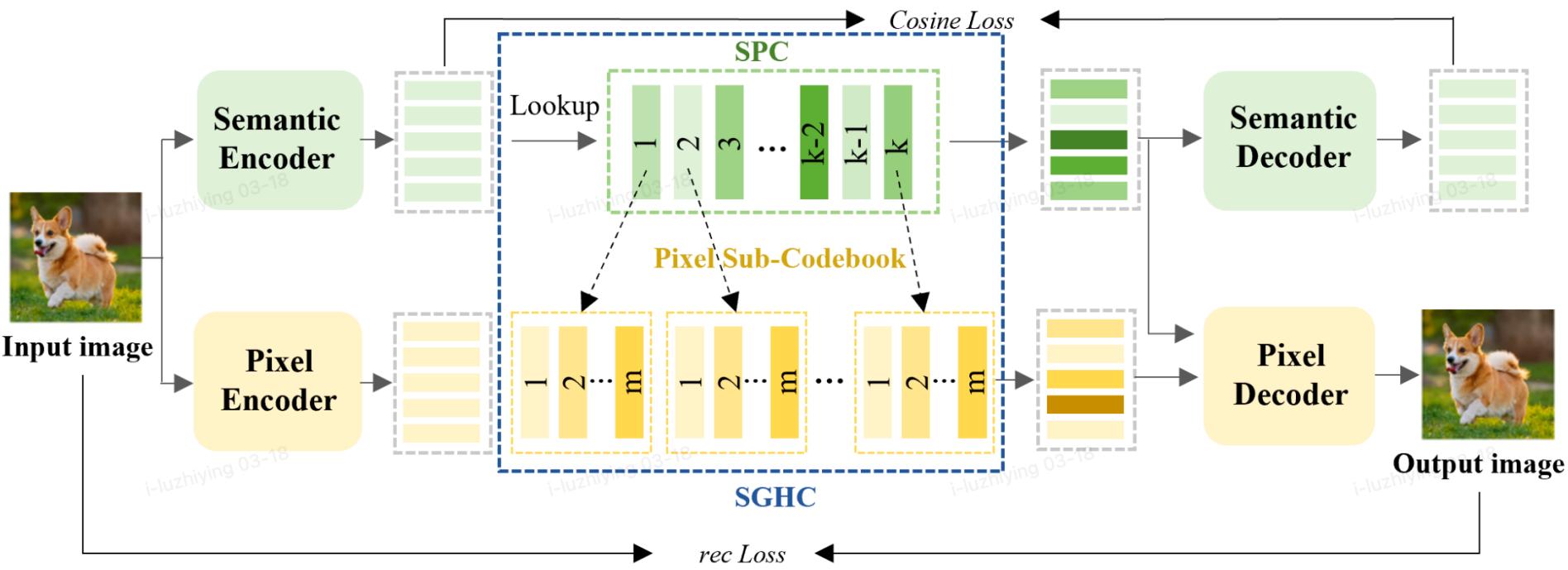


■ Special task tokens for distinguishing various tasks ■ Image tokens ■ Text tokens ■ Sequence with masked tokens

# 生成理解统一编码器

21

- 将语义信息融入到tokenizer中





- 作者介绍
- 背景介绍
- 本文方法
- 实验效果
- 总结反思

# Dual DiT

23

- 采用SD3的MMDiT模型，对于text采用mask diffusion方式实现，对于图像采用FM

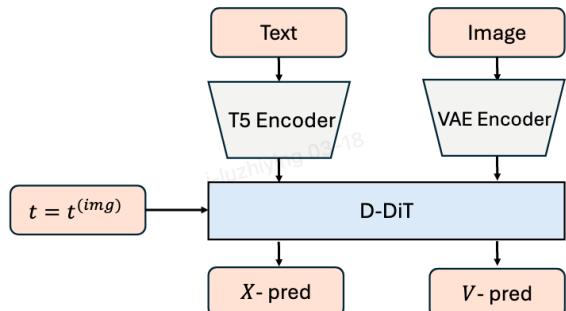
$$L_{\text{image}}$$

$$= \mathbb{E}_{t, q^{(\text{img})}} \left\| \mathbf{v}_\theta \left( \mathbf{x}_t^{(\text{img})}, t, \mathbf{x}^{(\text{txt})} \right) - (\boldsymbol{\epsilon} - \mathbf{x}^{(\text{img})}) \right\|_2^2,$$

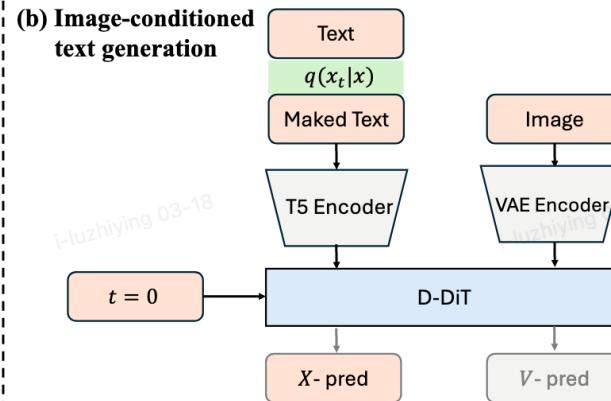
$$L_{\text{text}}$$

$$= \mathbb{E}_{q^{(\text{txt})}} \left[ -\frac{1}{K} \sum_{i=1}^K \log[\mathbf{x}_\theta(\mathbf{x}_{t_i}^{(\text{txt})}, \mathbf{x}^{(\text{img})}) \cdot \mathbf{x}] / t_i \right],$$

(a) Overview



(b) Image-conditioned text generation



(c) Text-conditioned image generation

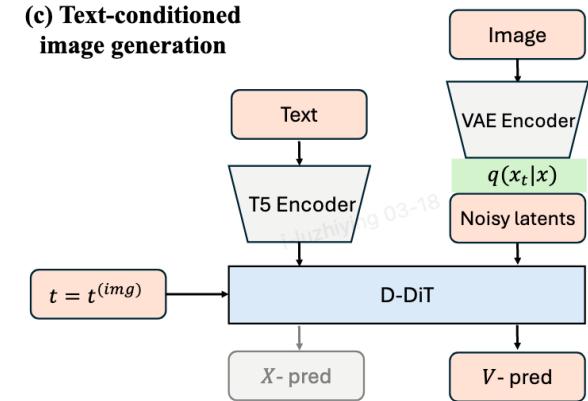


Figure 1. Our proposed model, the Dual Diffusion Transformer (D-DiT) that simultaneously models image and text distributions via a joint denoising diffusion training loss. **a)** An overview of the model architecture. The gray blocks (T5 encoder, image autoencoder) are kept fixed throughout training and inference. **b)** During training for (image-conditioned) text denoising, the text input is randomly masked while the image is noise-free. **c)** During training for text-conditioned image denoising, the image is randomly noised while the text is noise-free.



# Dual DiT

24

- 将文生图和图像理解视为条件概率
- 文生图结合CFG（类别引导生成收敛）

文生图

$$\mathbf{x} \sim p(\mathbf{x}^{(img)} | \mathbf{x}^{(txt)})$$

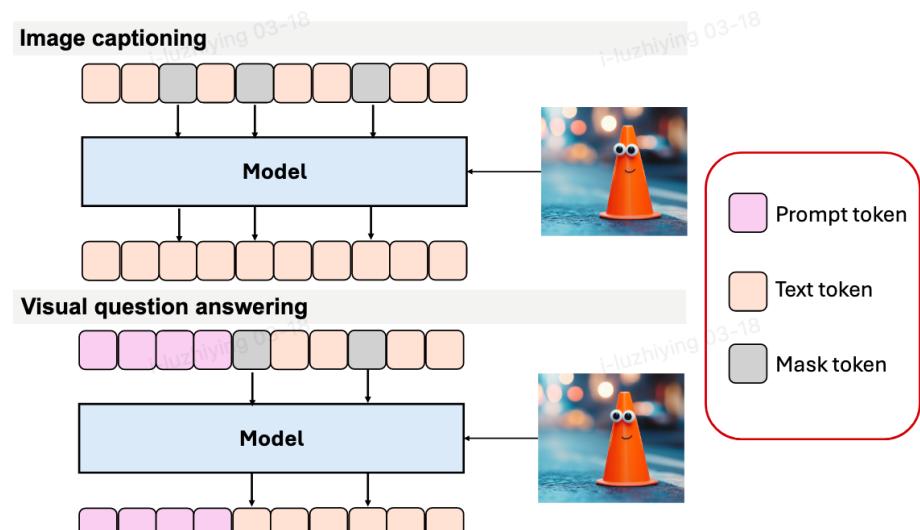
$$\tilde{v}_t = s\mathbf{v}_\theta \left( \mathbf{x}_t^{(img)}, t, \mathbf{x}^{(txt)} \right) + (1-s)\mathbf{v}_\theta \left( \mathbf{x}_t^{(img)}, t, \emptyset \right)$$

Caption

$$\mathbf{x} \approx \mathbf{x}_\theta(\mathbf{x}_t^{(txt)}, \mathbf{x}^{(img)}; t=0)$$

VQA

$$p(\mathbf{x}^{(answer)} | \mathbf{x}^{(img)}, \mathbf{x}^{(question)}).$$





- 作者介绍
- 背景介绍
- 本文方法
- 实验效果
- 总结反思



# 实验效果

26

Model	params (B)	Overall	Objects		Counting	Colors	Position	Color attribution
			Single	Two				
PixArt- $\alpha$ [10]	0.6	0.48	0.98	0.50	0.44	0.80	0.08	0.07
SD V2.1	0.9	0.50	0.98	0.51	0.44	0.85	0.07	0.17
DALL-E 2 [62]	6.5	0.52	0.94	0.66	0.49	0.77	0.10	0.19
SDXL [57]	0.9	0.55	0.98	0.74	0.39	0.85	0.15	0.23
DALL-E 3	-	0.67	0.96	0.87	0.47	0.83	0.43	0.45
CoDI [71]	-	0.31	0.89	0.16	0.16	0.65	0.02	0.01
LWM [48]	7	0.47	0.93	0.41	0.46	0.79	0.09	0.15
SEED-X [24]	17	0.49	0.97	0.58	0.26	0.80	0.19	0.14
Chameleon [72]	7	0.39	-	-	-	-	-	-
Show-O [80]	1.3	0.68	0.98	0.80	0.66	0.84	0.31	0.50
Transfusion [85]	8	0.67	-	-	-	-	-	-
SD3 [19]	2	0.62	0.98	0.74	0.63	0.67	0.34	0.36
D-DiT (ours)	2	0.65	0.97	0.80	0.54	0.76	0.32	0.50



# 实验效果

27

Model	Params	Text # trainable	Image Backbone	MS-COCO CIDEr ↑	VQAv2 Acc. ↑	VizWiz Acc. ↑	OKVQA Acc. ↑	MME Acc. ↑	GQA Acc. ↑	POPE Acc. ↑
LLaVA-1.5 [46]	13B	AR	-	-	81.8	57.5	-	1500.1	64.7	86.4
BLIP-2 [39]	13B	AR	-	-	65.0	19.6	-	1293.8	41.0	85.5
IDEFICS [36]	9B	AR	-	-	50.9	-	-	-	-	-
QWEN-VL [6]	7B	AR	-	-	78.2	38.9	-	1487.5	57.5	-
OpenFlamingo [5]	9B	AR	-	65.5	43.5	-	-	-	-	-
Flamingo [2]	9B	AR	-	79.4	51.8	28.8	44.7	-	-	-
CM3Leon [83]	7B	AR	AR	61.6	47.6	37.6	23.8	-	-	-
Chameleon [72]	7B	AR	AR	18.0	-	-	-	-	-	-
LWM [49]	7B	AR	AR	-	55.8	11.6	-	-	44.8	75.2
Show-O (256×256) [80]	1.3B	AR	Diffusion	-	64.7	-	-	1014.9	54.2	76.2
Show-O (512×512) [80]	1.3B	AR	Diffusion	-	69.4	-	-	1097.2	58.0	80.0
Transfusion [85]	7B	AR	Diffusion	29.0	-	-	-	-	-	-
D-DiT (Ours, 256×256)	2B	Diffusion	Diffusion	-	59.5	19.4	28.5	897.5	55.1	79.2
D-DiT (Ours, 512×512)	2B	Diffusion	Diffusion	56.2	60.1	29.9	25.3	1124.7	59.2	84.0

Task	$T = 4$	8	16	32	64	128
VQAV2 (acc.)	58.8	58.0	59.3	60.5	60.0	59.6
MS-COCO (CIDEr)	20.2	35.3	46.5	51.3	56.2	54.5

# 实验

28



t=1

**Q:** Provide a brief description of the given image. **A:** [MASK] image [MASK] a serene and beautiful [MASK] scene taken [MASK] [MASK] [MASK] [MASK] [MASK], which appears to be a [MASK] [MASK] [MASK]. From [MASK] wooden balcony [MASK] [MASK] of a clear [MASK] [MASK] where the snowy town [MASK] the base [MASK] [MASK] horizon. [MASK] [MASK] [MASK] ray [MASK] illuminate the landscape [MASK] casting [MASK] a warm glow [MASK] the azure backdrop [MASK]</s></s></s> [MASK] [MASK]</s></s></s></s></s></s></s></s></s></s></s></s></s></s> [MASK]</s></s> [MASK] [MASK]</s> [MASK] [MASK]</s></s></s></s> [MASK]</s></s> [MASK] [MASK]</s></s></s></s></s> [MASK]

**Q:** Provide a brief description of the given image. **A:** [MASK] image [MASK] a serene and beautiful [MASK] scene taken [MASK] [MASK] [MASK] [MASK] [MASK], which appears to be a [MASK] [MASK] [MASK]. From [MASK] wooden balcony [MASK] [MASK] of a clear [MASK] [MASK] where the snowy town [MASK] the base [MASK] [MASK] horizon. [MASK] [MASK]' [MASK] [MASK] ray [MASK] illuminate the landscape [MASK] casting [MASK] a warm glow [MASK] the azure backdrop [MASK]</s></s></s> [MASK] [MASK]</s></s></s> [MASK]</s></s></s></s></s></s> [MASK]</s></s></s> [MASK] [MASK]</s></s> [MASK] [MASK]</s></s> [MASK]</s></s></s></s> [MASK]</s></s> [MASK]</s></s></s></s> [MASK]

**Q:** Provide a brief description of the given image. **A:** The image presents a serene and beautiful winter scene taken from a vantage point, which appears to be a mountain range. From the wooden balcony in front of a clear blue sky where the snowy town at the base meets the horizon. The sun's rays illuminate the landscape, casting a warm glow against the azure backdrop.

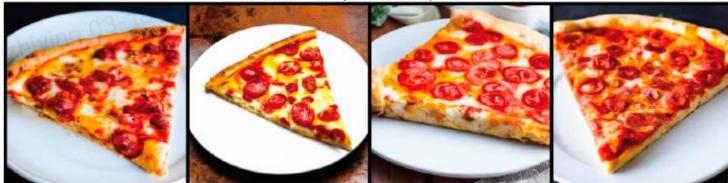
t=0



# 实验效果

29

A slice of pizza on a plate.



A bouquet of roses in a vase.



A flock of birds flying across the sky.



A train crossing a bridge over a river.



A child blowing bubbles in a backyard.



A carousel in motion at an amusement park.



A city skyline silhouetted against a sunrise.



A ballerina dancing on a stage under spotlight.



A medieval castle atop a hill surrounded by fog.



An explorer discovering ancient ruins in a jungle.



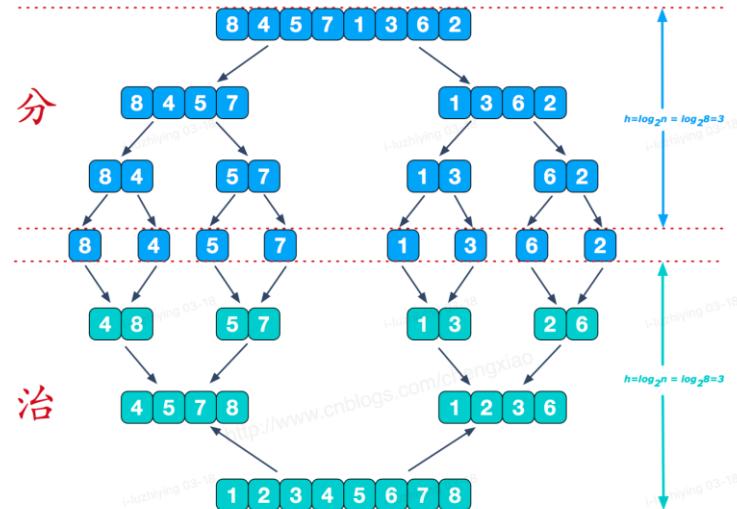


- 作者介绍
- 背景介绍
- 本文方法
- 实验效果
- 总结反思

# 总结反思

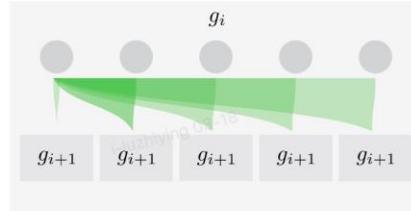
31

- 递归比遍历更加逼近智能
- 自回归式-空间上遍历，扩散式-空间上递归
- 自回归，优点：符合知识先验，缺点：不具备单点优化能力
- 扩散，优点：单点优化增强，缺点：不具备空间外推性
- **自回归+递归=AGI?**

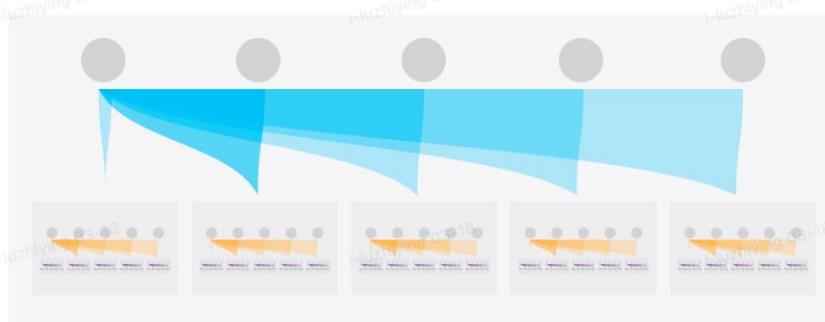


## Fractal Generative Models

Tianhong Li<sup>1</sup> Qinyi Sun<sup>1</sup> Lijie Fan<sup>2</sup> Kaiming He<sup>1</sup>



(a) generator



(b) fractal from the generator



# 谢谢！