Text Diffusion Models, Multimodal LLMs

COMP7607 — Lecture 9

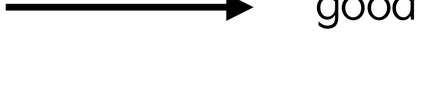
Lingpeng Kong

Department of Computer Science, The University of Hong Kong

Generate a sentence

$$p(\boldsymbol{x}) = \prod_i p(x_i|\boldsymbol{x}_{<i})$$

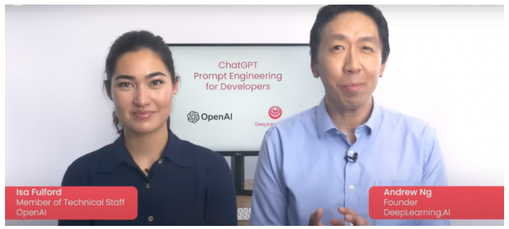
Don’t just believe that it is because something is trendy that it is …



good

LLMs as Agents

LLM Reasoning



Prompt Engineering

Non-autogressive Text Generation

Tokyo is the largest city in the world

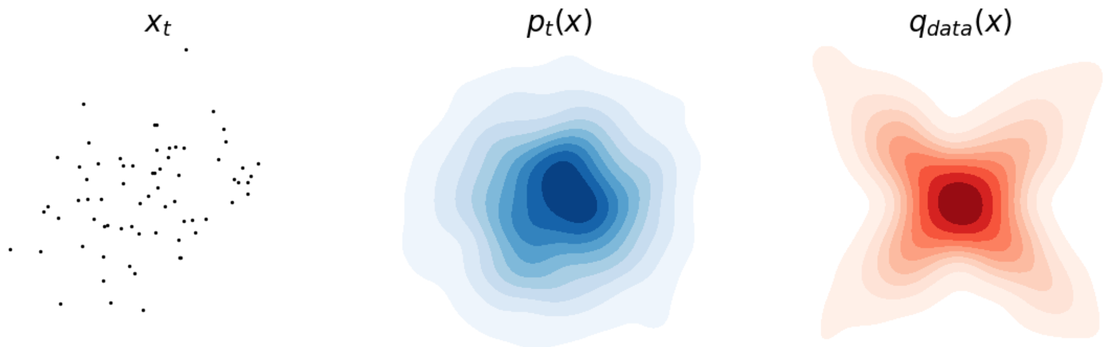
↑ Iterative Refinement

Shanghai is the largest city in the world

4+10 14 14-12 2 13*2=24
3*13-39,9+10-19,39-19-24
1+10-11 13-11=2,7*2=24

⇧ Parallel Generation

〔mask〕〔mask〕〔mask〕〔mask〕〔mask〕〔mask〕〔mask〕〔mask〕



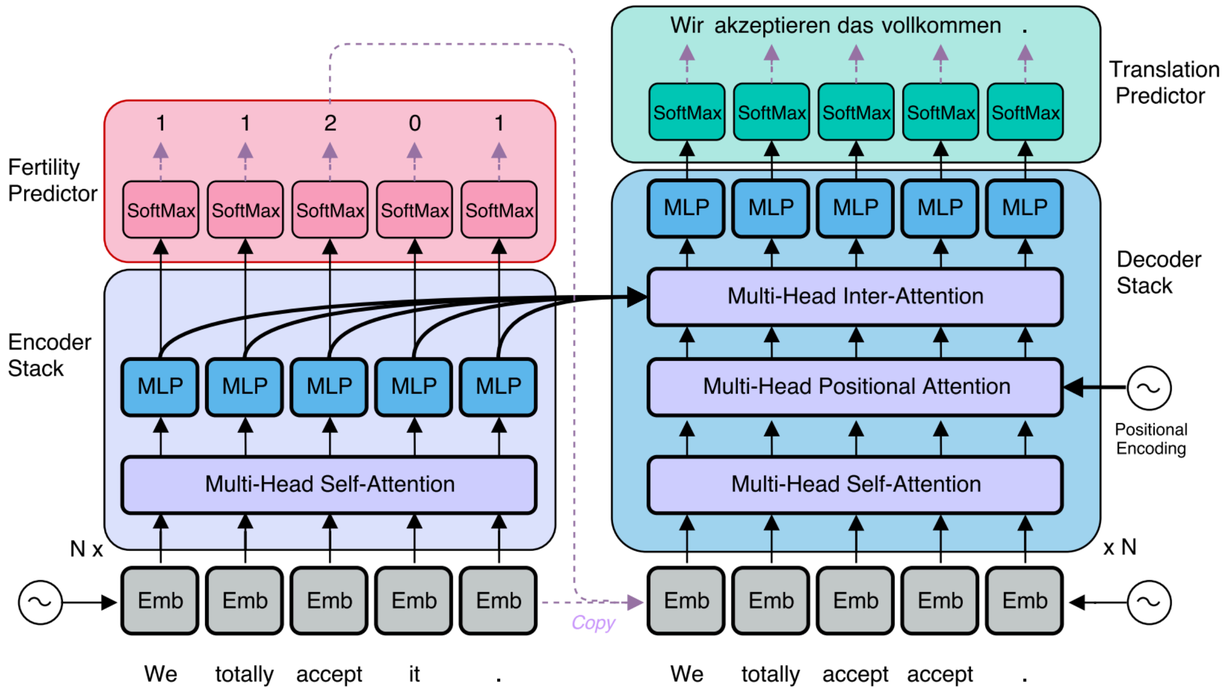
In a different “family” of p(x) than the autoregressive ones

Non-autogressive Neural Machine Translation

Models	WMT14 En→De	WMT14 De→En
NAT	17.35	20.62
NAT (+FT)	17.69	21.47
NAT (+FT + NPD $s = 10$)	18.66	22.41
NAT (+FT + NPD $s = 100$)	19.17	23.20
Autoregressive ($b = 1$)	22.71	26.39
Autoregressive ($b = 4$)	23.45	27.02

"Modality Conflicts"

many thanks
thank you ⟶ thank thanks

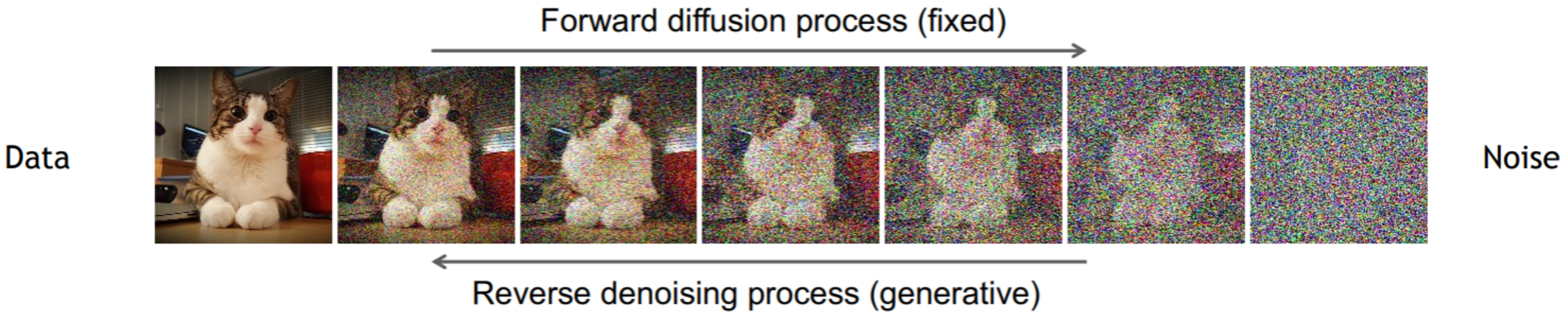


Jiatao Gu, et al., Non-autogressive Ne

Diffusion Models

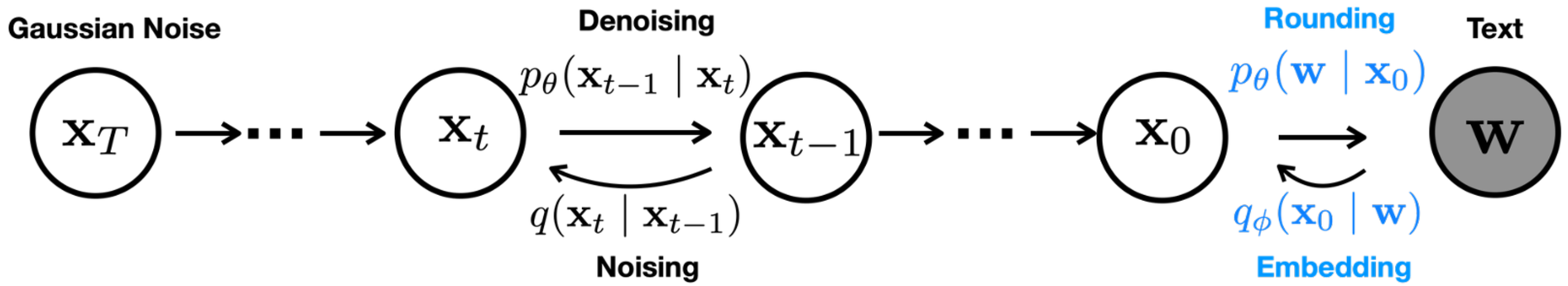
Forward-backward formulation

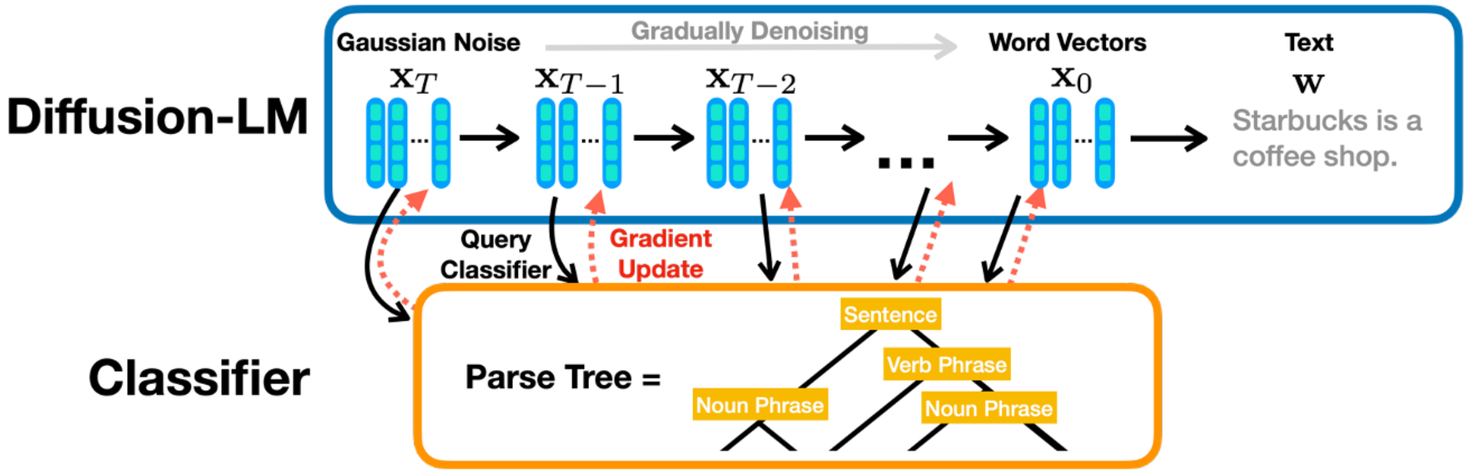
The forward process gradually injects noise to the input
The backward process denoises to recover the original data



"Creating noise from data is easy; creating data from noise is generative modeling." (Song et al., 2021)

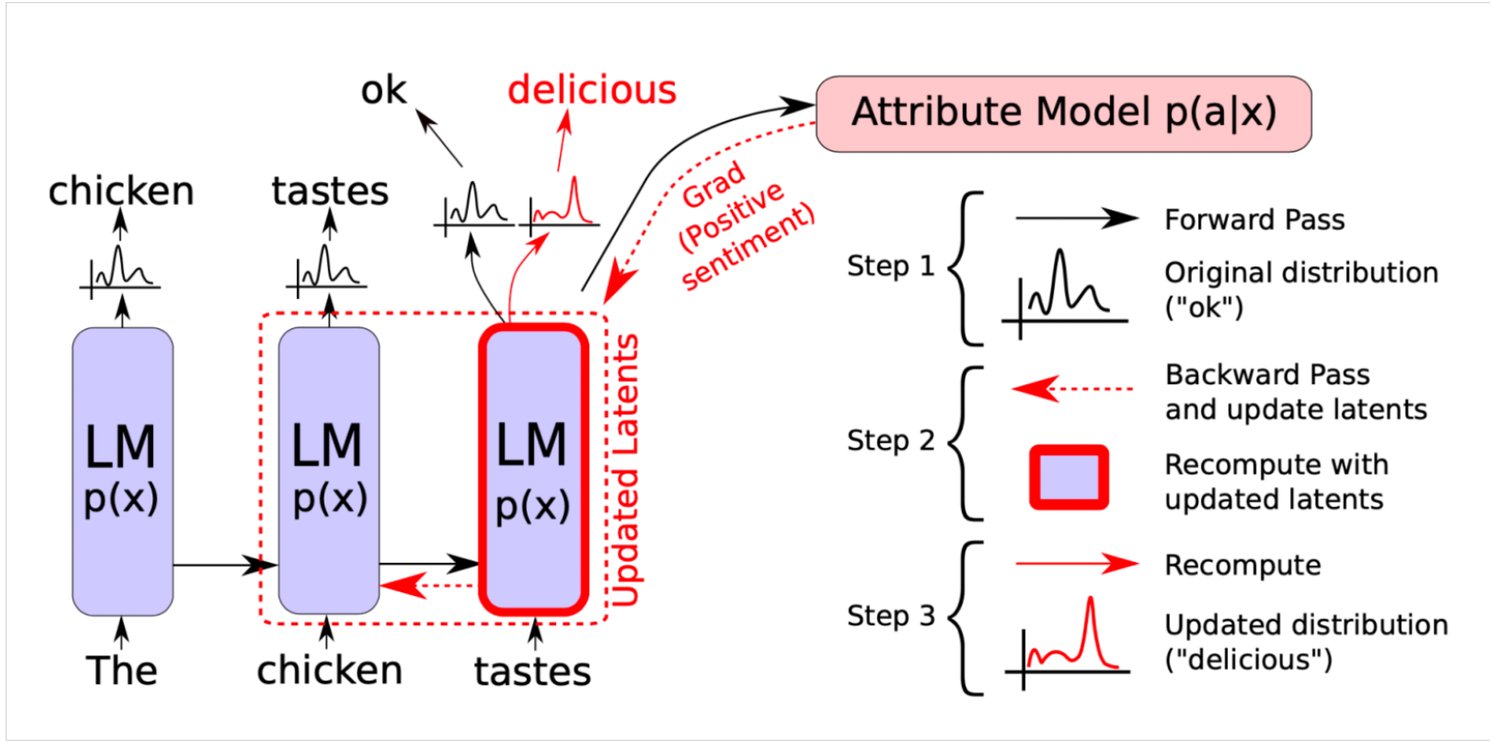
Text Diffusion Models

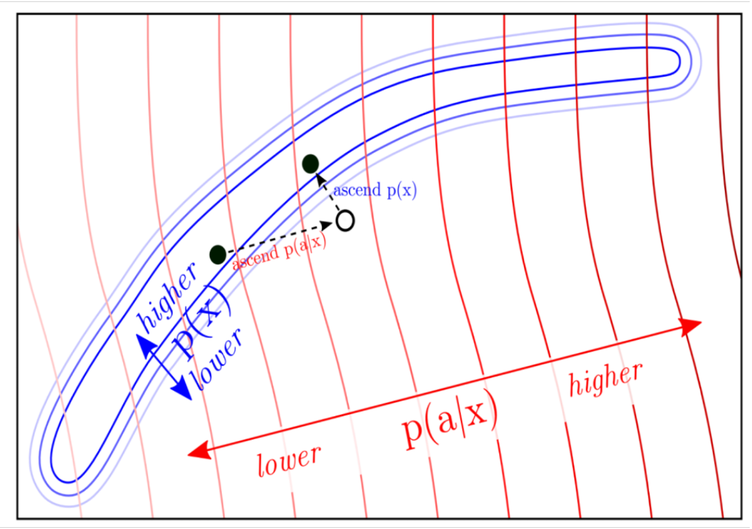




Xiang Lisa Li, et at., Diffusion-LM Improves Controllable Text Generation, NeurIPS 2022

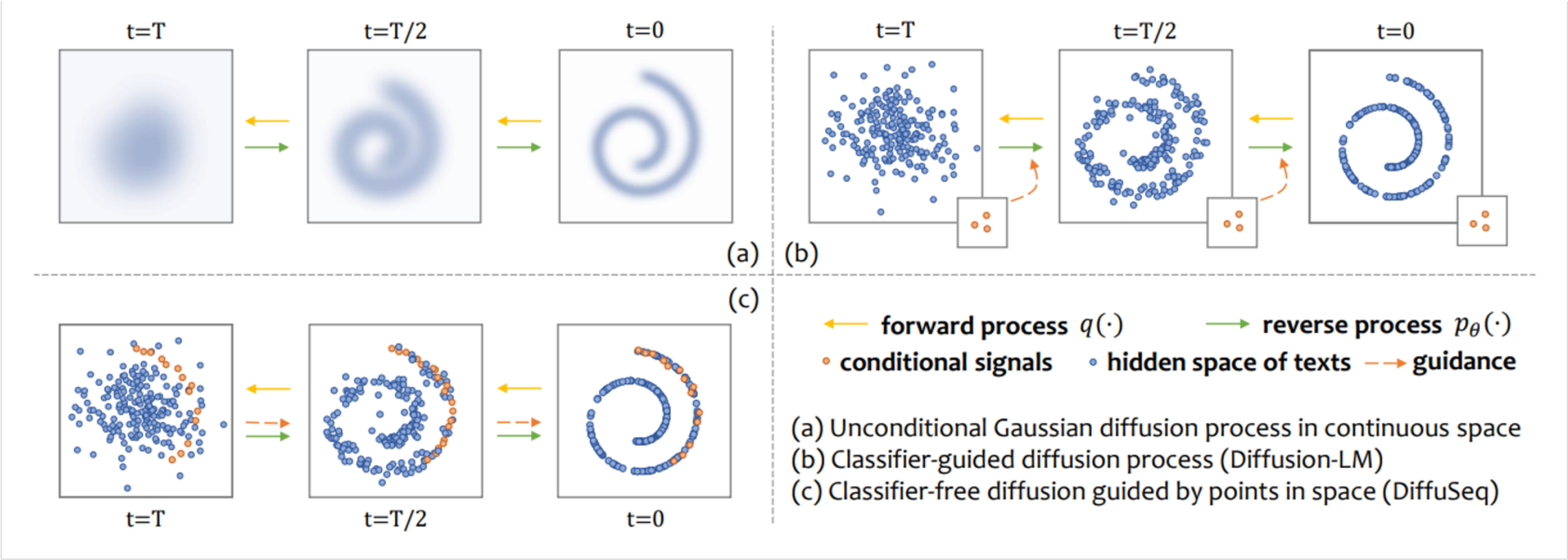
Classifier Guidance





Sumanth Dathathri, et at., Plug and Play Language Models: A Simple Approach to Controlled Text Generation, ICLR 2020

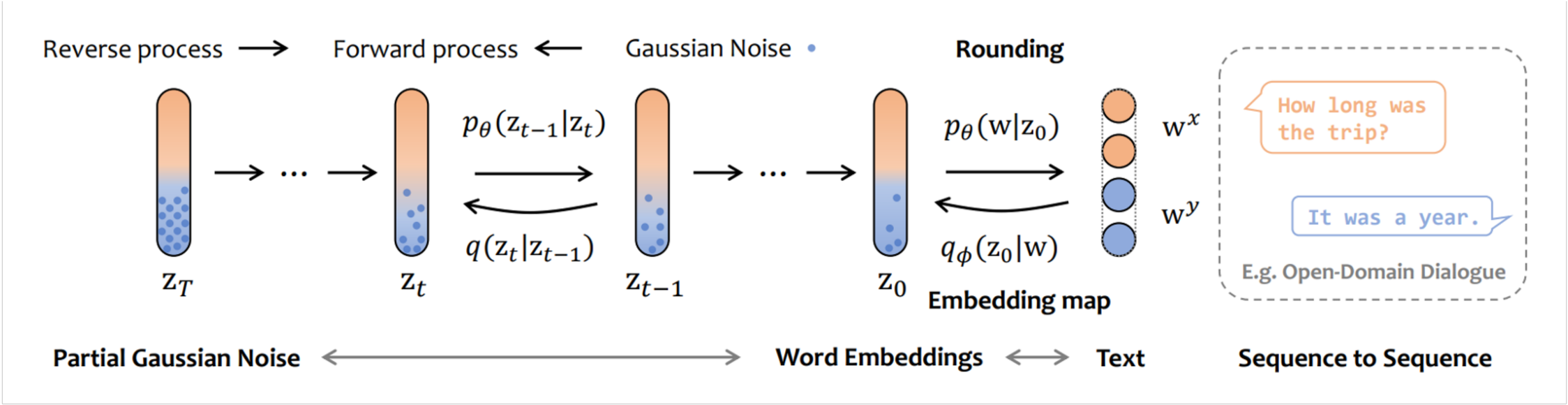
Sequence to Sequence Text Diffusion



"Seq2Seq" tasks: **x→y**

Shansan Gong, et at., DiffuSeq: sequence to sequence text generation with diffusion models, ICLR 2023

Sequence to Sequence Text Diffusion



$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_{q(\mathbf{z}_{1:T}|\mathbf{z}_0)}\Bigg[\underbrace{\log\frac{q(\mathbf{z}_T|\mathbf{z}_0)}{p_\theta(\mathbf{z}_T)}}_{\mathcal{L}_T} + \sum_{t=2}^{T}\underbrace{\log\frac{q(\mathbf{z}_{t-1}|\mathbf{z}_0,\mathbf{z}_t)}{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}}_{\mathcal{L}_{t-1}} + \underbrace{\log\frac{q_\phi(\mathbf{z}_0|\mathbf{w}^{x\oplus y})}{p_\theta(\mathbf{z}_0|\mathbf{z}_1)}}_{\mathcal{L}_0} - \underbrace{\log p_\theta(\mathbf{w}^{x\oplus y}|\mathbf{z}_0)}_{\mathcal{L}_{\text{round}}}\Bigg].$$

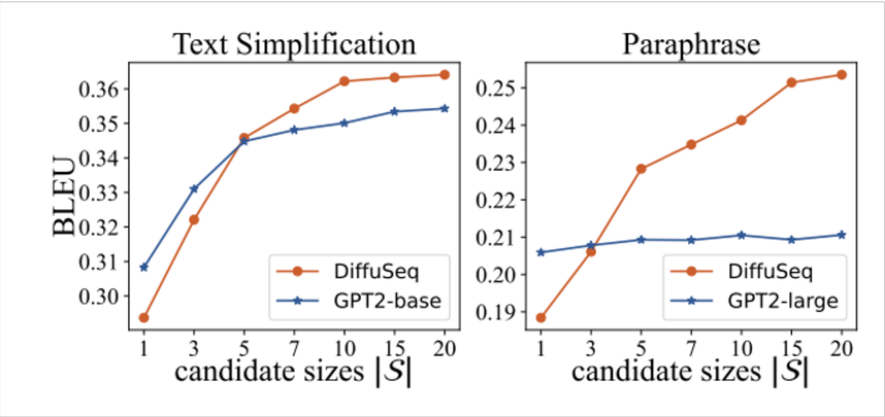
x

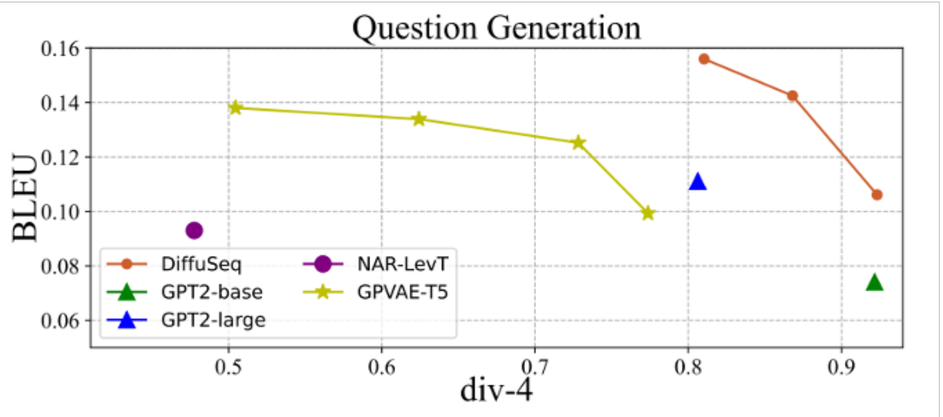
$$\min_\theta \mathcal{L}_{\text{VLB}} = \min_\theta \Bigg[\sum_{t=2}^{T}||\mathbf{y}_0 - \tilde{f}_\theta(\mathbf{z}_t,t)||^2 + ||\text{EMB}(\mathbf{w}^y) - \tilde{f}_\theta(\mathbf{z}_1,1)||^2 + \mathcal{R}(||\mathbf{z}_0||^2)\Bigg]$$

Shansan Gong, et at., DiffuSeq: sequence to sequence text generation with diffusion models, ICLR 2023

Sequence to Sequence Text Diffusion

Diversity Ensures Quality



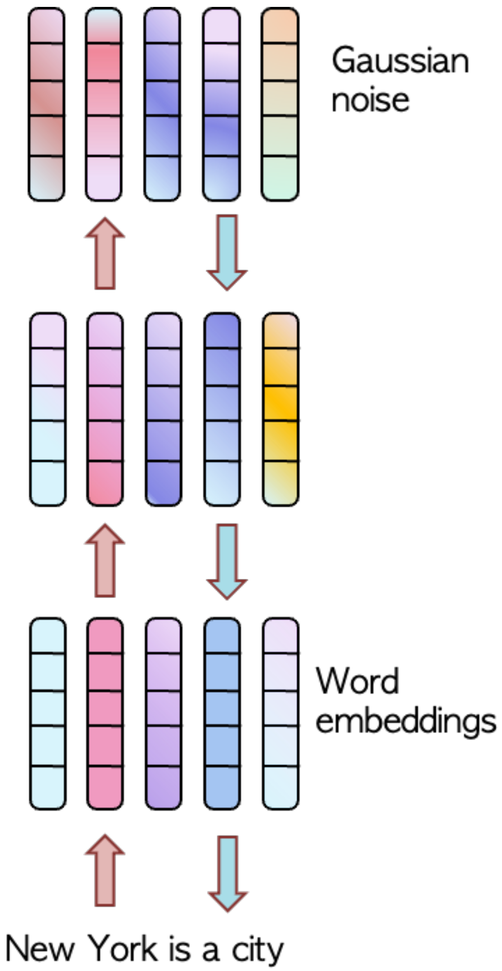


Statement: *The Japanese yen is the official and only currency recognized in Japan.*
Question: *What is the Japanese currency?*

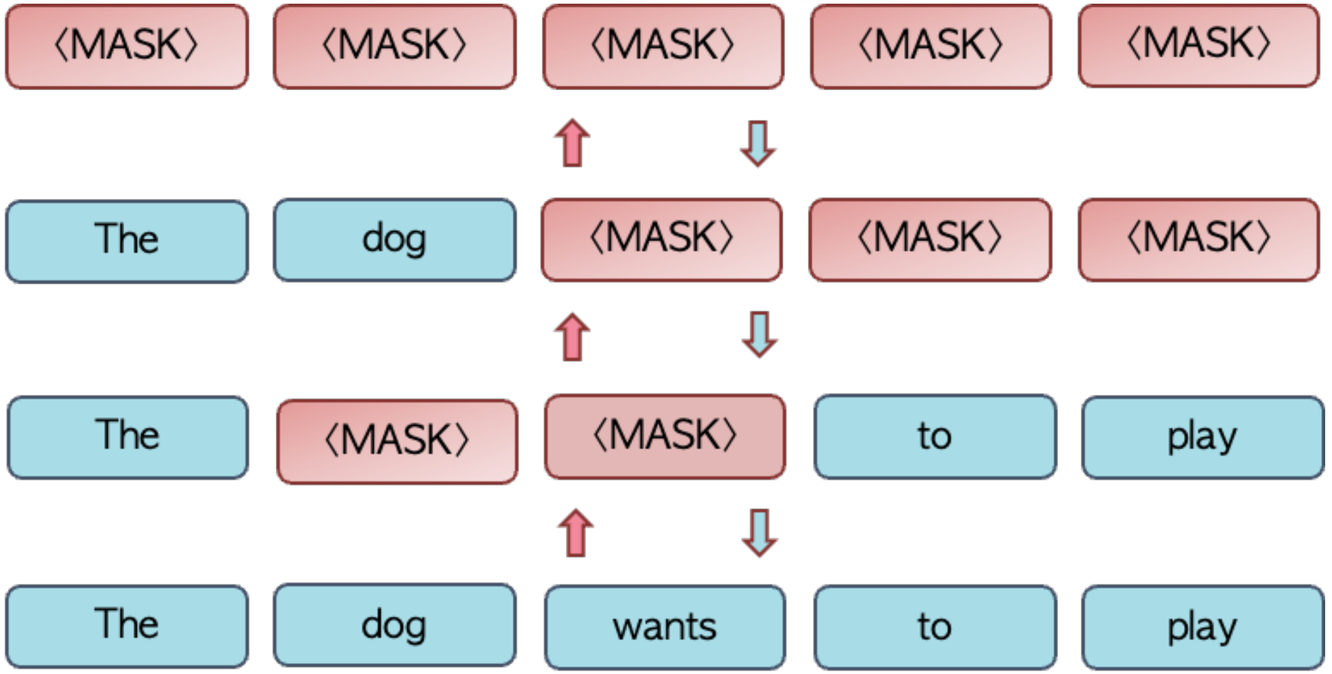
GPVAE-T5	**NAR-LevT**
* What is the japanese currency	* What is the basic unit of currency for Japan ?
* What is the japanese currency	* What is the basic unit of currency for Japan ?
* What is the japanese currency	* What is the basic unit of currency for Japan ?
GPT2-large finetune	**DiffuSeq**
* What is the basic unit of currency for Japan?	* What is the Japanese currency
* What is the Japanese currency	* Which country uses the “yen yen” in currency
* What is the basic unit of currency for Japan?	* What is the basic unit of currency?

Shansan Gong, et at., DiffuSeq: sequence to sequence text generation with diffusion models, ICLR 2023

Text Diffusion Models



Continuous Diffusion



Discrete Diffusion

Discrete Diffusion - Forward Process

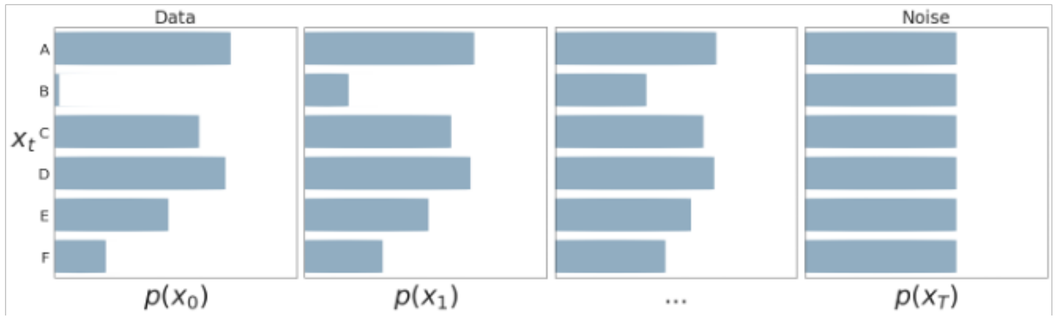
$$q(x_0) = p_{\text{data}}(x)$$

Noise: $q(x_t \mid x_{t-1}) = \beta_t x_{t-1} + (1 - \beta_t) q_{\text{noise}}$

Enjoys closed-form "marginal" distribution:

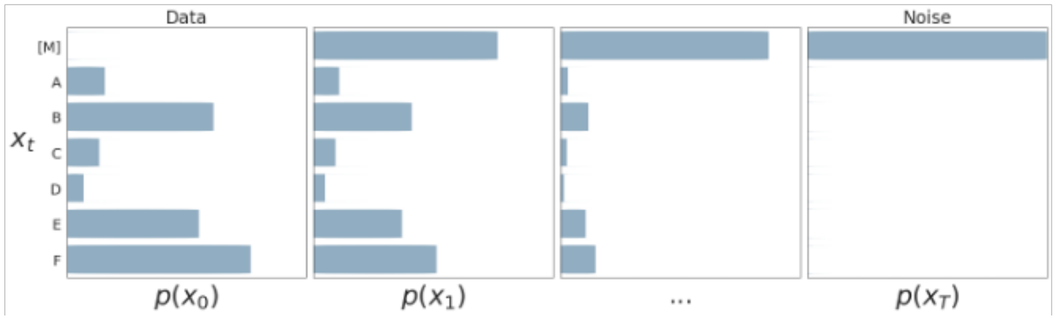
$$q(\mathbf{x}_t|\mathbf{x}_0) = \alpha_t \mathbf{x}_{t-1} + (1 - \alpha_t) q_{\text{noise}}, \quad \alpha_t := \prod_{i=1}^{t} \beta_i$$

"Multinomial"



$$q_{\text{noise}} = \frac{1}{K}$$

"Absorbing"



$$q_{\text{noise}} = \delta([\mathrm{M}])$$

Discrete Diffusion - Learning

Goal: $p_\theta(x_{t-1} \mid x_t) \approx q(x_{t-1}|x_t)$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_0) \geq \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T|\mathbf{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}_0,\mathbf{x}_1,\ldots,\mathbf{x}_T)}{q(\mathbf{x}_1,\ldots,\mathbf{x}_T|\mathbf{x}_0)}\right]$$

$$= \mathbb{E}_q\left[\underbrace{\log p_{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_1)}_{\mathcal{L}_1(\boldsymbol{\theta})} - \sum_{t=2}^{T}\underbrace{\mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \parallel p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{\mathcal{L}_t(\boldsymbol{\theta})}\right] + \text{const.}$$

Enjoys closed-form "marginal" distribution:

"Multinomial" $$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \frac{\left(\beta_t\mathbf{x}_t + (1-\beta_t)\frac{\mathbf{1}}{K}\right) \odot \left(\alpha_{t-1}\mathbf{x}_0 + (1-\alpha_{t-1})\frac{\mathbf{1}}{K}\right)}{\mathbf{x}_t^\top\left(\alpha_t\mathbf{x}_0 + (1-\alpha_t)\frac{\mathbf{1}}{K}\right)}$$

"Absorbing" $$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \begin{cases}\mathbf{x}_t, & \text{if } \mathbf{x}_t \neq \mathbf{e}_{[M]},\\ \frac{1-\alpha_{t-1}}{1-\alpha_t}\mathbf{e}_{[M]} + \frac{\alpha_{t-1}-\alpha_t}{1-\alpha_t}\mathbf{x}_0, & \text{if } \mathbf{x}_t = \mathbf{e}_{[M]}\end{cases}$$

Sampling Reparameterization

Reveals a latent routing mechanism

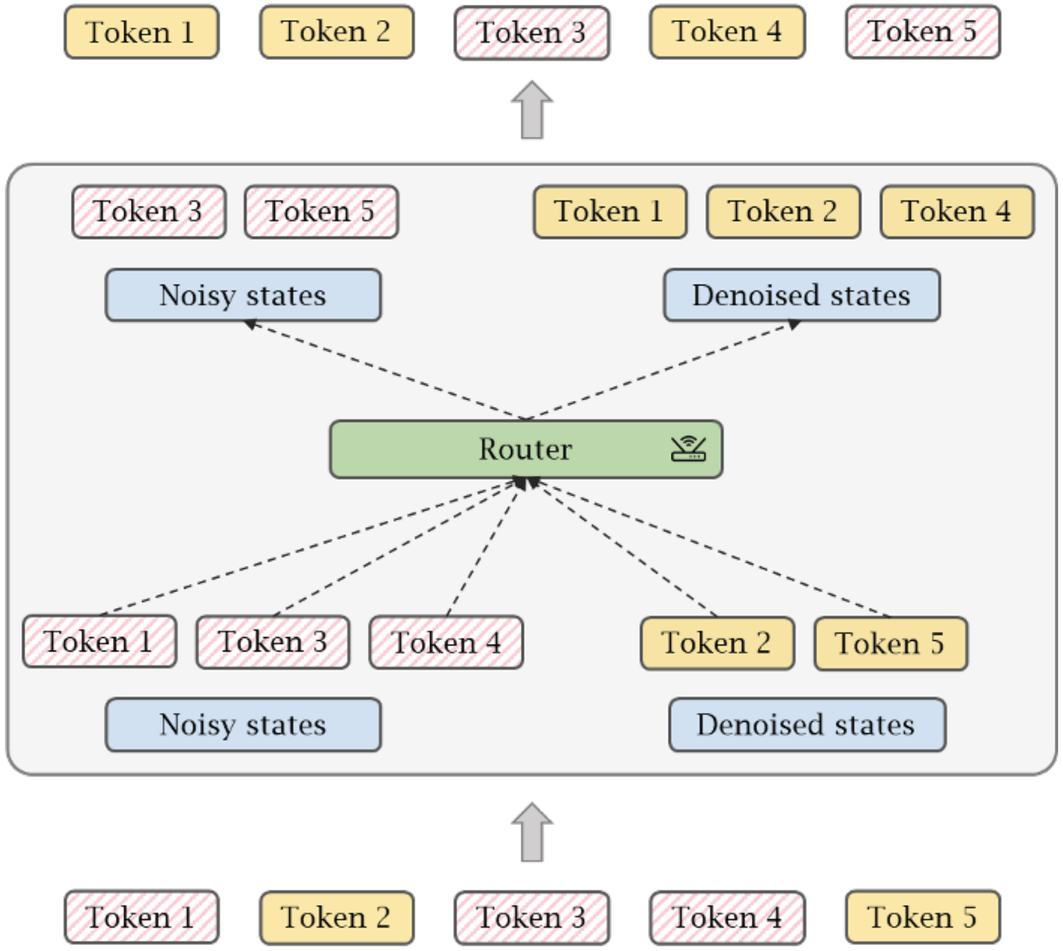
$$b_t = \mathbf{1}_{\mathbf{x}_t=\mathbf{x}_0}$$
$$v_{t-1}^{(1)} \sim \text{Bernoulli}\left(\lambda_{t-1}^{(1)}\right), \quad \mathbf{u}_t^{(1)} \sim q_{\text{noise}}$$
$$v_{t-1}^{(2)} \sim \text{Bernoulli}\left(\lambda_{t-1}^{(2)}\right), \quad \mathbf{u}_t^{(2)} \sim q_{\text{noise}}(\mathbf{x}_t)$$
$$\mathbf{x}_{t-1} = b_t \left[v_{t-1}^{(1)}\mathbf{x}_t + \left(1 - v_{t-1}^{(1)}\right)\mathbf{u}_t^{(1)}\right] + (1 - b_t)\left[v_{t-1}^{(2)}\mathbf{x}_0 + \left(1 - v_{t-1}^{(2)}\right)\mathbf{u}_t^{(2)}\right]$$

Sequence x_{t-1}

Route with $v_{t-1} \coloneqq [v_{t-1}^{(1)}, v_{t-1}^{(2)}]$

Calculate $b_t = \mathbf{1}_{x_t=x_0}$

Sequence x_t



Lin Zheng et at., A Reparameterized Discrete Diffusion Model for Text Generation, 2023

Training RDMs

Trained with ELBO:

$$\mathcal{L}_t := -\lambda_{t-1}^{(2)} \sum_{n=1}^{N} \mathbf{1}_{\mathbf{x}_{t,n} \neq \mathbf{x}_{0,n}} \mathbf{x}_{0,n}^{\top} \log f\left(\mathbf{x}_{t,n}; \boldsymbol{\theta}\right)$$

Time-dependent weighting

Evaluated at noisy tokens

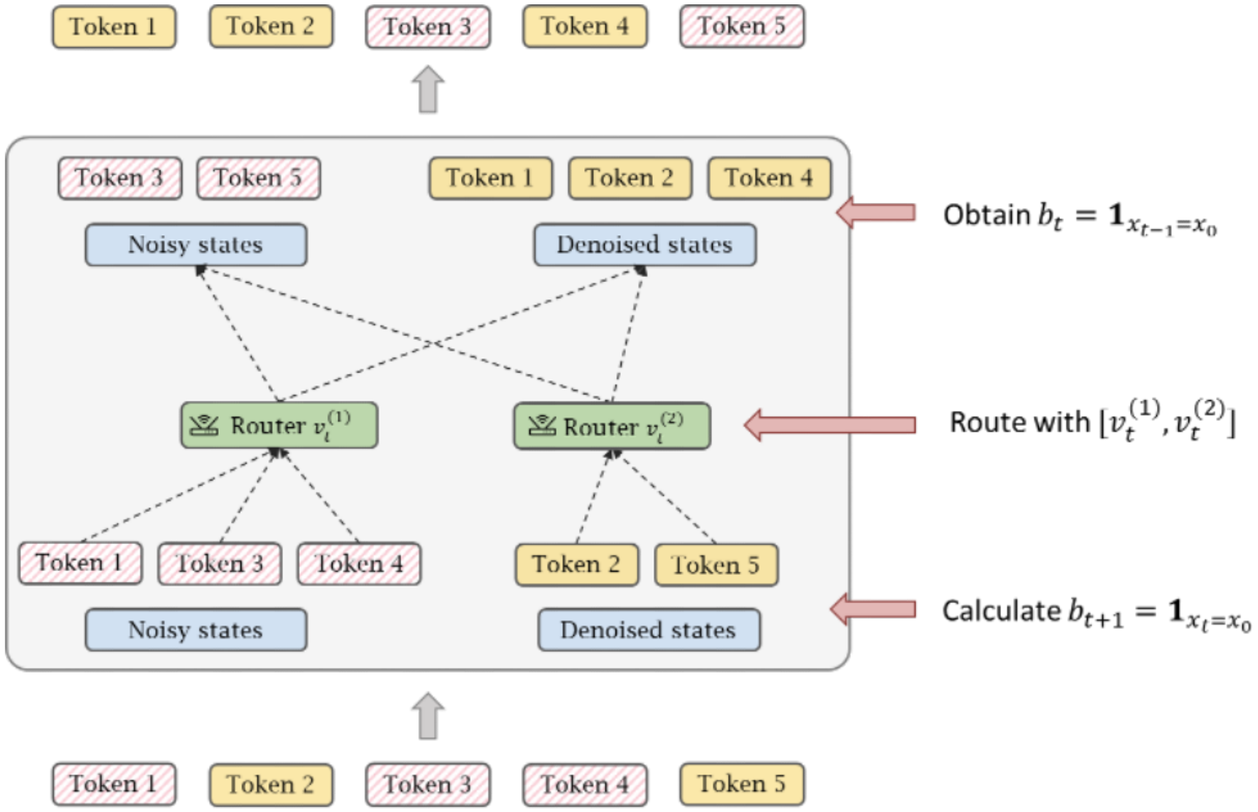
Simple Cross-Entropy Loss

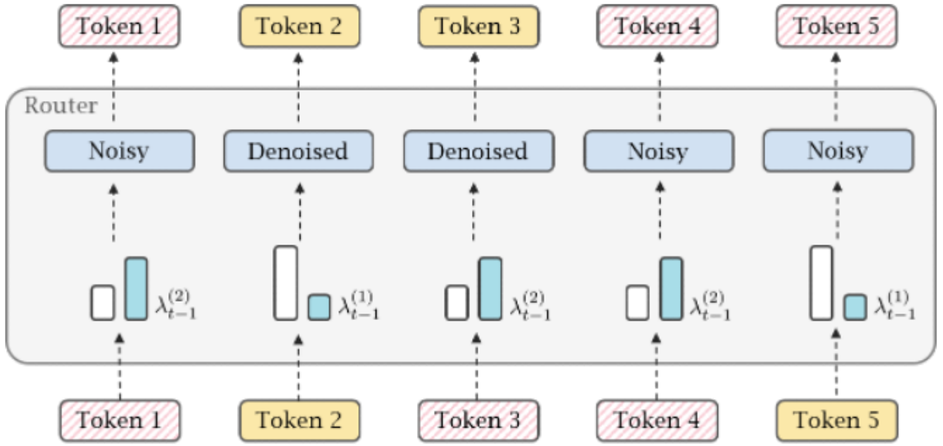
$$\lambda_{t-1}^{(2)} = 1 - \frac{t-1}{T}$$

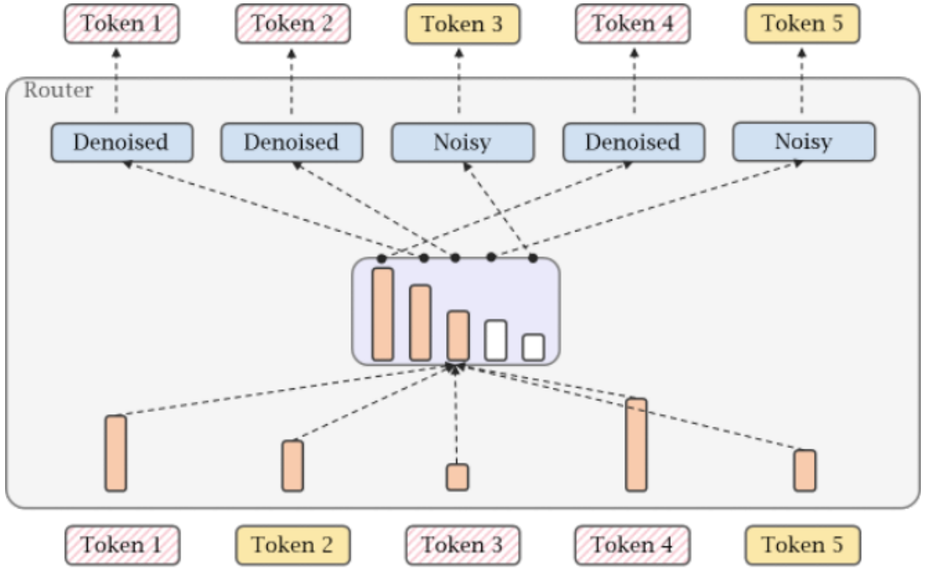
Router-agnostic

Lin Zheng et at., A Reparameterized Discrete Diffusion Model for Text Generation, 2023

Decoding RDMs







Confident-based Router

Lin Zheng et at., A Reparameterized Discrete Diffusion Model for Text Generation, 2023

Experimental Results of RDMs

Close the gap with autoregressive models (Machine Translation)

	Model	# Iterations	IWSLT14 DE-EN		WMT16 EN-RO		WMT14 EN-DE	
			Vanilla	Reparam.	Vanilla	Reparam.	Vanilla	Reparam.
Continuous Diffusion	CDCD (Dieleman et al., 2022)	200	–		–		20.0*	
Discrete Diffusion	Multinomial Diffusion (Hoogeboom et al., 2021)	2	23.05	28.01	26.61	30.16	4.28	21.43
		4	24.24	30.57	27.81	31.70	4.31	24.05
		10	21.28	32.23	25.25	33.00	6.94	25.63
		16	20.59	32.58	24.36	33.11	6.07	25.64
		25	20.06	32.84	23.94	33.31	3.69	26.04
	Absorbing Diffusion (Austin et al., 2021)	2	25.24	27.60	27.24	30.72	16.46	21.00
		4	26.93	31.47	29.16	32.60	19.48	24.26
		10	28.32	33.91	30.41	33.38	21.62	26.96
		16	28.38	34.41	30.79	33.82	22.07	27.58
		25	28.93	34.49	30.56	33.99	22.52	**27.59**
Auto-regressive Models	Transformer-base (Vaswani et al., 2017)	n.a.	**34.51**		**34.16**		27.53	

Lin Zheng et at., A Reparameterized Discrete Diffusion Model for Text Generation, 2023

Experimental Results of RDMs

Fix previously made errors in where vanilla discrete diffusion models get stuck

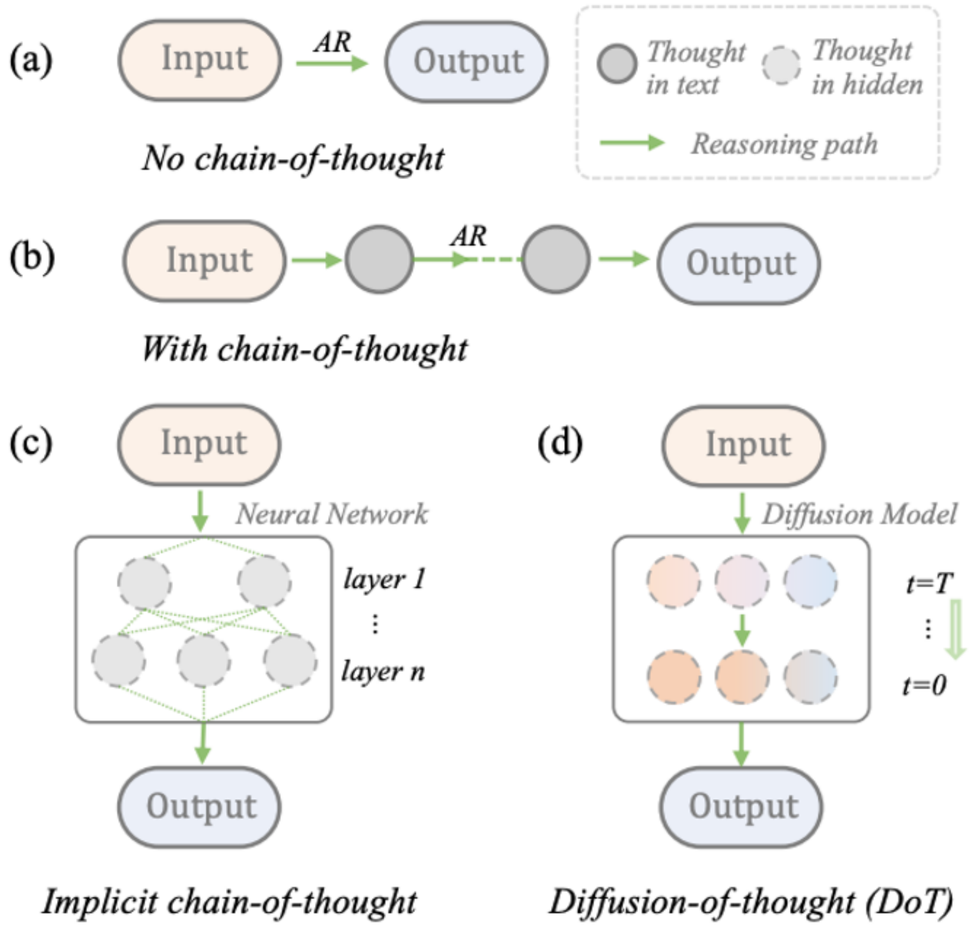
Source: i have only 2 months for my ca cpt exams how do i prepare?
Reference: i want to crack ca cpt in 2 months. how should i study?

	# Iter.	Decodes
Absorbing	0	∘ <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M>
	1	∘ <M> <M> <M> <M> <M> ca <M> ca <M> <M> ca <M> <M> <M> <M>
	2	∘ <M> <M> <M> <M> <M> ca <M> ca <M> <M> ca <M> <M> <M> <M>
	3	∘ <M> <M> <M> <M> <M> ca - ca cp <M> ca <M> <M> exam <M>
	4	∘ <M> can <M> prepare <M> ca - ca cp <M> ca <M> ##t exam ?
	5	∘ how can i prepare for ca - ca cp ##t ca cp ##t exam ?
RDM-absorbing	0	∘ <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> <M>
	1	∘ how <M> i <M> <M> <M> <M> <M> <M> <M> <M> <M> <M> ?
	2	∘ how <M> i prepare for ca <M> ##t <M> <M> <M> <M> <M> ?
	3	∘ how <M> i prepare for ca cp ##t <M> <M> <M> months months ?
	4	∘ how <M> i prepare for ca cp ##t exam in two months <M> ?
	5	∘ how can i prepare for ca cp ##t exam in two months left ?

Lin Zheng et at., A Reparameterized Discrete Diffusion Model for Text Generation, 2023

Autoregressive LLMs, what else?

Diffusion of Thoughts (DoTs)



Diffusion of Thoughts

🔍 Questions:
When Freda cooks canned tomatoes into sauce, they lose half their volume. Each 16 ounce can of tomatoes that she uses contains three tomatoes. Freda's last batch of tomato sauce made 32 ounces of sauce.
How many tomatoes did Freda use?

🤖 CoT Decoding:
<

ʚ DoT Decoding:
<<32/16=2>> <<12/12=2*3 ## *2=2= ######

ʚɞ MP-DoT Decoding:
<<32===2>>

Jiacheng Ye, Shansan Gong, Liheng Chen et at., Diffusion of Thoughts: Chain-of-Thought Reasoning in Diffusion Language Models, 2024

Mutilmodal LLMs

- Motivation: Why do we need a versatile large multimodal models (LMMs)?

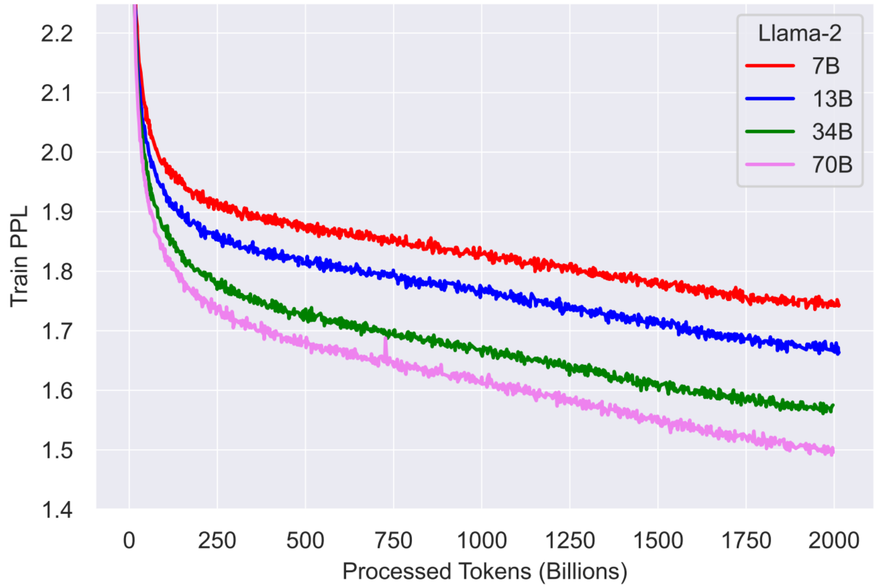
- Progress: What recipe do we have for building capable LMMs?
  - Architecture
  - Data
  - Training Recipe

Future Directions: What we can do in the area of LMMs?

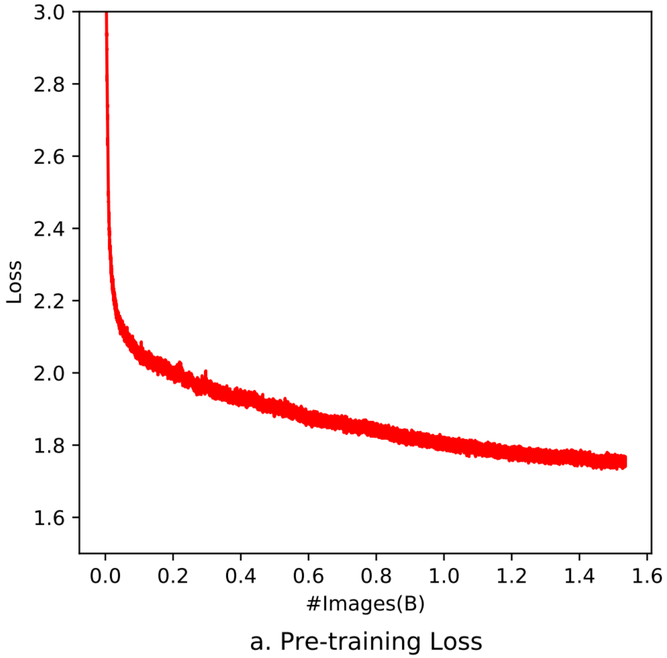
Slides for Multimodal LLMs from Lei Li (https://lilei-nlp.github.io/), xhs ID: tobiaslee

Motivations: Why LLMs

Compression of AGI



LLaMa-2 Training Loss



a. Pre-training Loss

Qwen-VL-Chat (7B)

Image = 256 tokens
Caption ~= 64 tokens

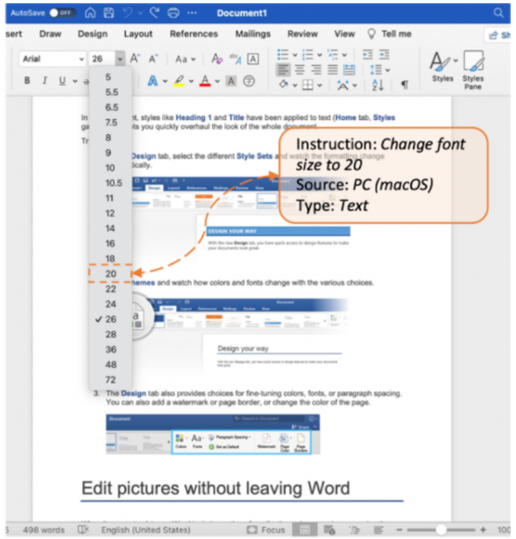
~ 500B tokens -> ~1.8

LLaMA-2-7B 500B tokens -> 1.9

Motivations: Why LLMs

Real-world applications are beyond text.

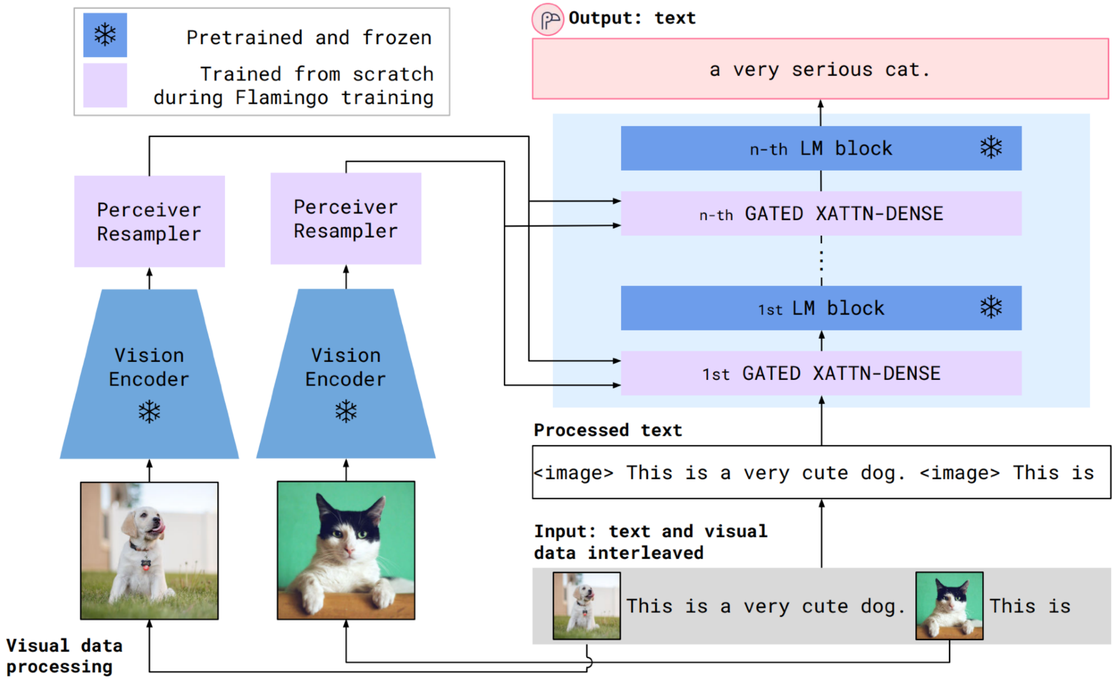
Embodied Robot



GUI Agent

RT-2: New model translates vision and language into action
SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents

Architecture Overview



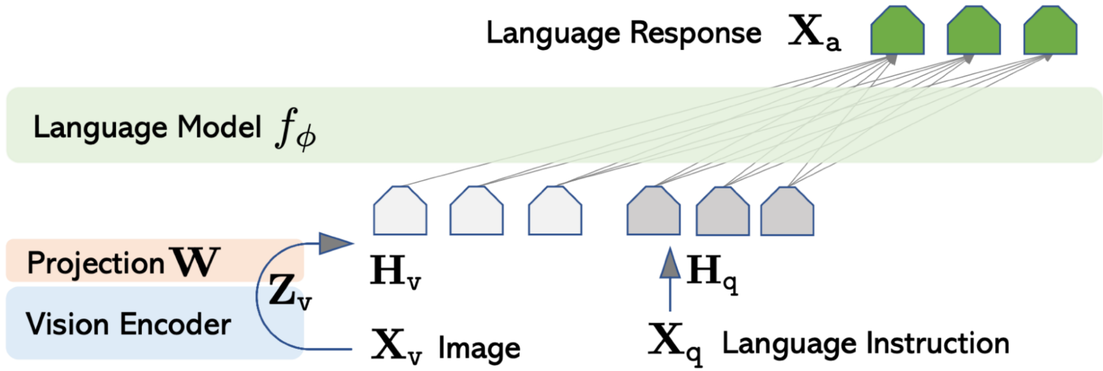


Figure 1: LLaVA network architecture.

Flamingo
Images are integrated via
Gated Cross-Attention

LLaVA / Qwen-VL-Chat / MiniGPT4 ..
Images are treated as Word Embeddings
More prevailing now

Architecture Design Space

- Large Language Model:
  - Use the best LLM available.
- Vision Encoder:
  - Influence of the image resolution.
  - Which vision encoder shall we use?
- Modality Projector:
  - How to **effectively** represent the image in word embedding space?

How might GPT-4v deal with high resolution?

Here are some examples demonstrating the above.

- A 1024 × 1024 square image in `detail: high` mode costs 765 tokens
  - 1024 is less than 2048, so there is no initial resize.
  - The shortest side is 1024, so we scale the image down to 768 × 768.
  - 4 512px square tiles are needed to represent the image, so the final token cost is `170 * 4 + 85 = 765`.
- A 2048 × 4096 image in `detail: high` mode costs 1105 tokens
  - We scale down the image to 1024 × 2048 to fit within the 2048 square.
  - The shortest side is 1024, so we further scale down to 768 × 1536.
  - 6 512px tiles are needed, so the final token cost is `170 * 6 + 85 = 1105`.
- A 4096 × 8192 image in `detail: low` most costs 85 tokens
  - Regardless of input size, low detail images are a fixed cost.

GPT-4V Document

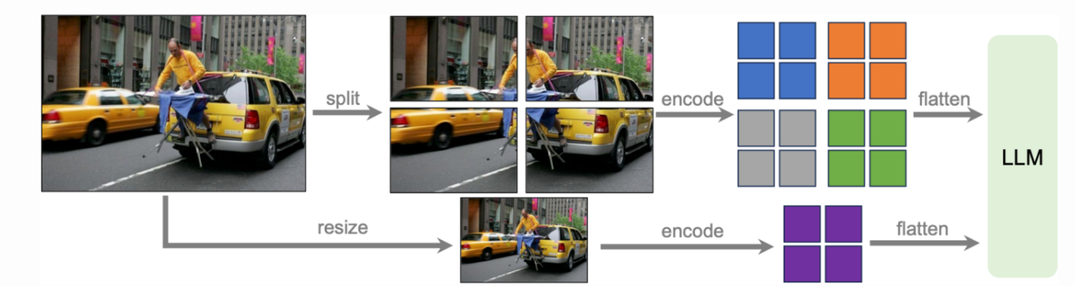
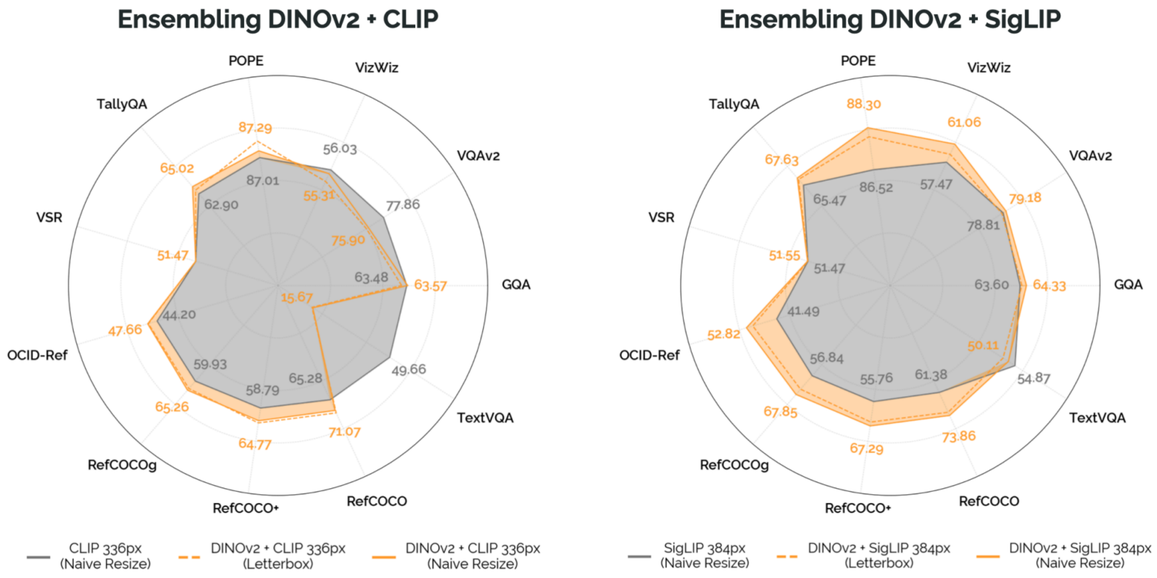


Illustration of dynamic high resolution scheme: a grid configuration of 2 × 2

LLaVA-Next cuts image into grids to reduce the computation overhead

LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

Vision Feature Types



GPT-4V Document

- Single CLIP Feautre is not comprehensive for grounding related tasks.
- Combination of SigCLIP and DINOv2 (image SSL) leads to the best results across benchmarks.

Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models

Modality Projector

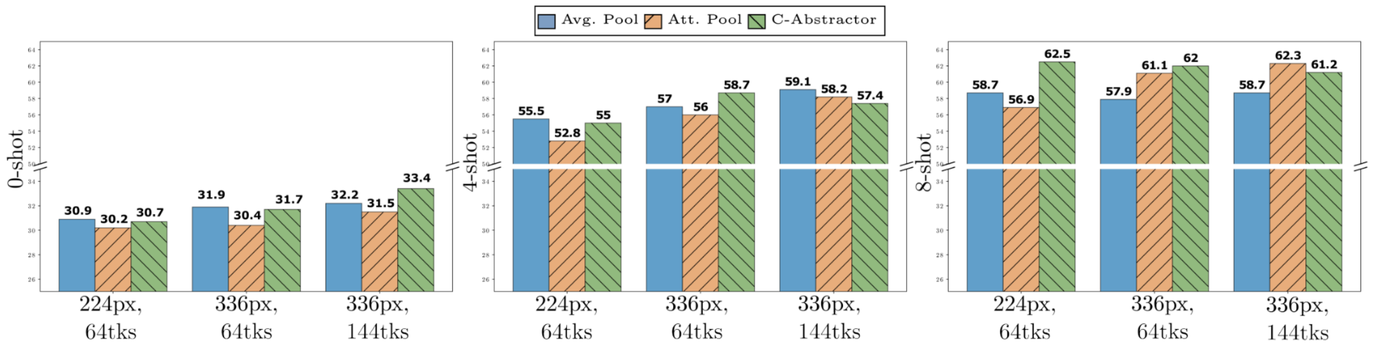
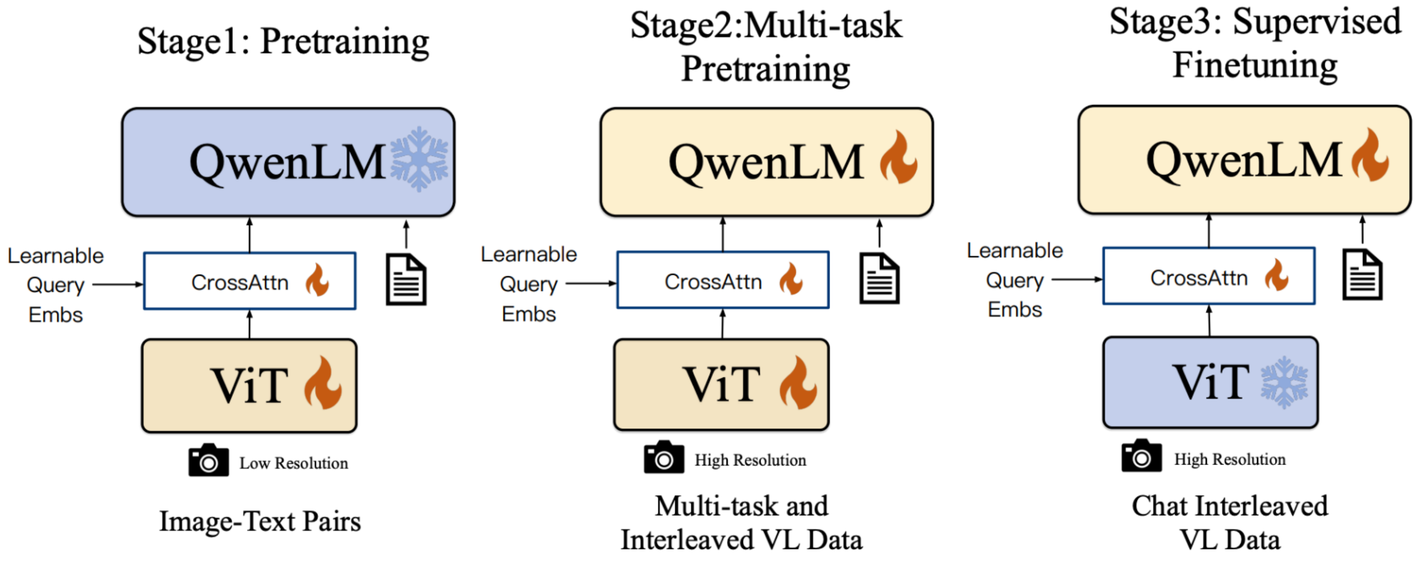


Fig. 4: 0-shot, 4-shot, and 8-shot ablations across different visual-language connectors for two image resolutions, and two image token sizes.

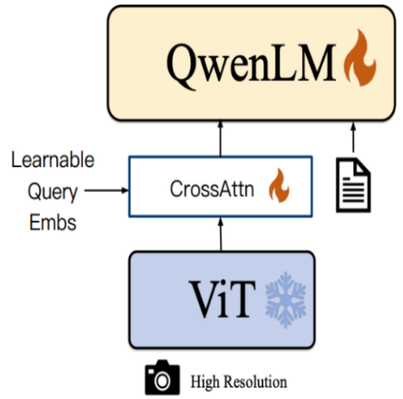
- Projector can be utilized to compress the visual tokens as well, which one shall we use?
- Avg Pooling v.s. Attn Pooling v.s C-Abstractor (ResNet Pooling)
- Findings:
  - Number of visual tokens and image resolution matters most, while the type of VL connector has little effect.

MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

Data & Training Stage Overview



Stage 4: Alignment according to Feedback

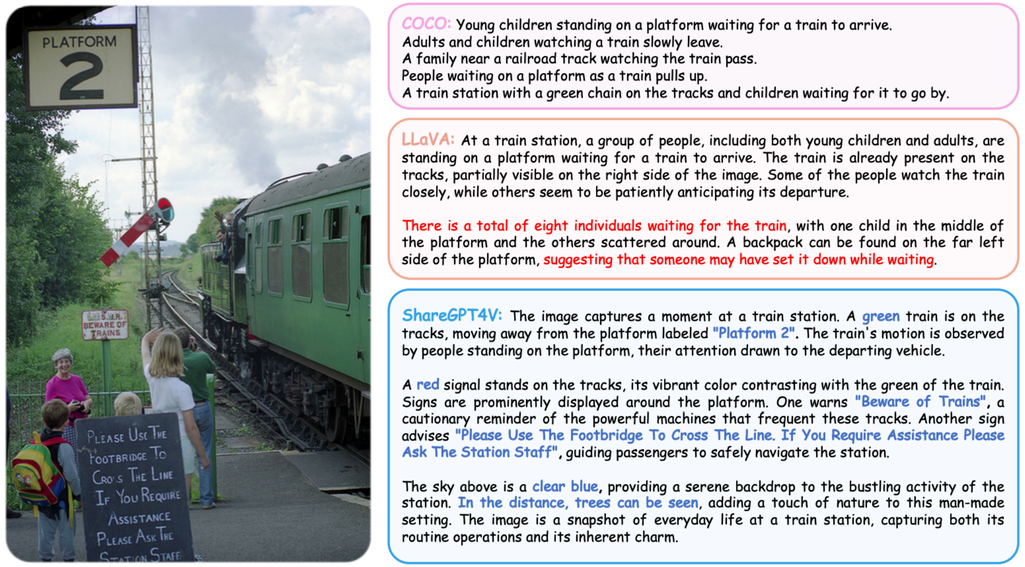


VL Feedback data

General VL Datasets

- Stage 1: Large-scale Paired dataset for modality alignment
  - Scale & Quality
  - Raw datasets: Laion5B, Datacomp-1B
  - LLM (LMMs) rewrites dataset: Laion-COCO, Capsfusion, ShareGPT4V

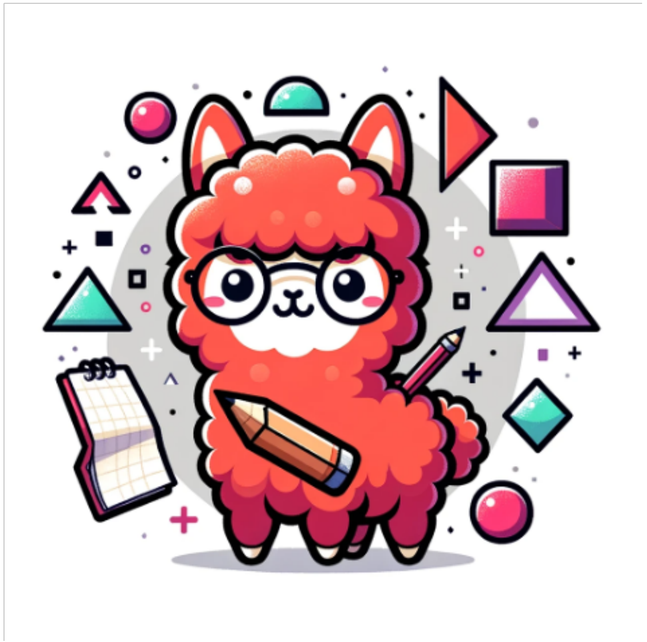
- Stage 2 & 3: Instruction tuning datasets / Chatty Dataset
  - M3IT (Academic VL tasks, https://m3-it.github.io), wide coverage.
  - LLaVA-Instruct (GPT-4/ChatGPT generated psuedo-mutlimodal dataset)
  - Specialized Domain Datasets: AI2D, OCR-VQA, LLaVAR, ChartVQA, G-LLaVA

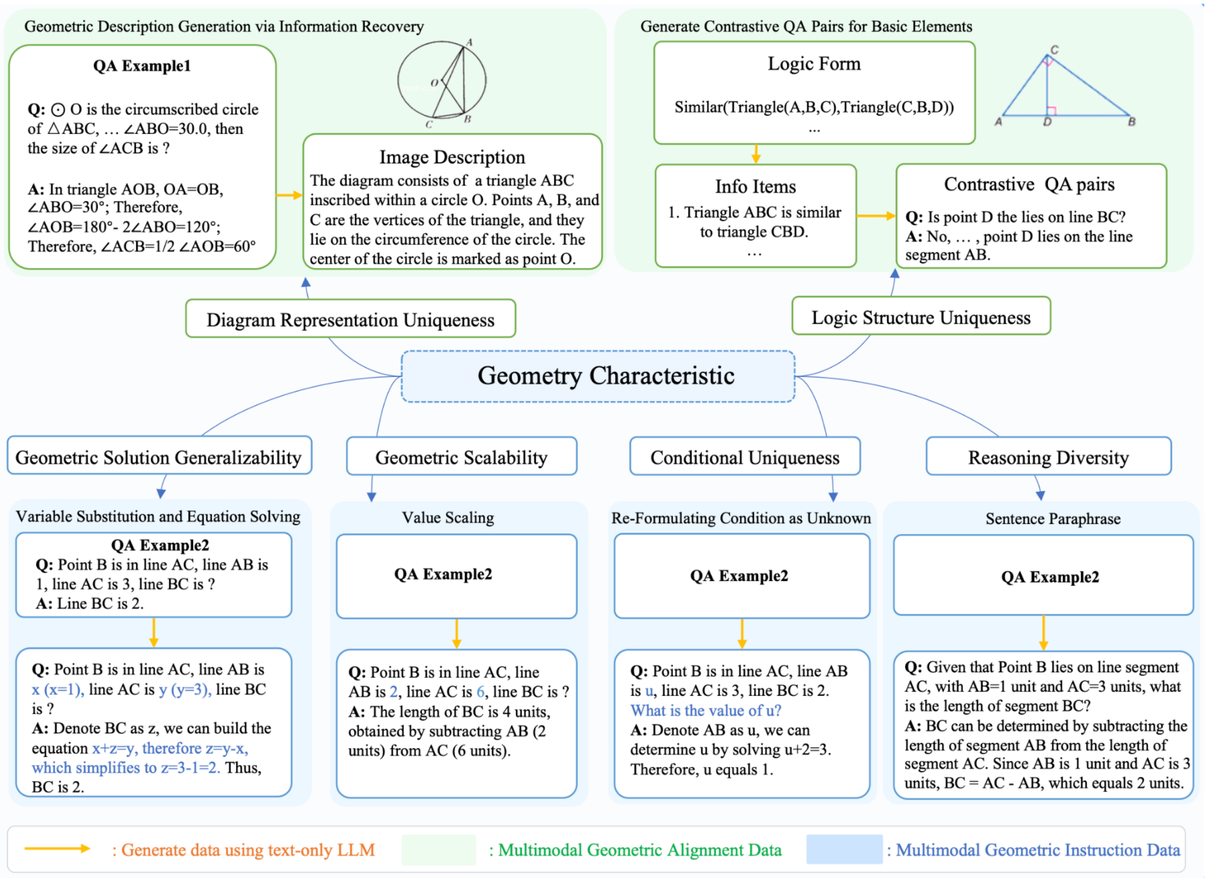


(a) Comparison of Captions' Quality

ShareGPT4V: Improving Large Multi-Modal Models with Better Captions

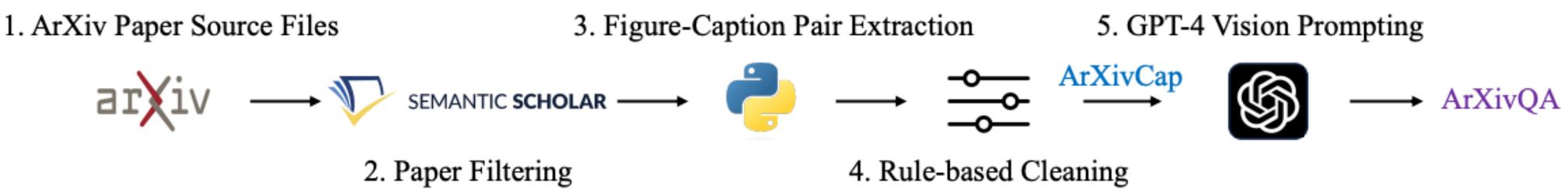
Datasets for LLMs Math Reasoning





G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model

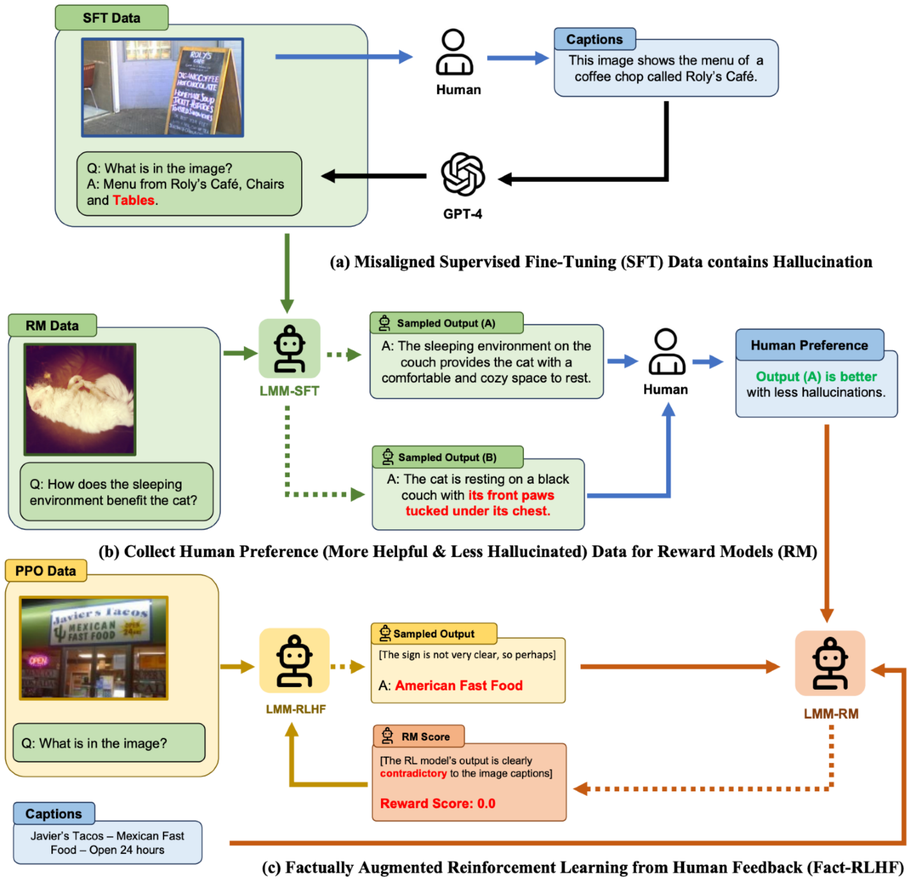
Multimodal ArXiv for LLMs Scientific Understanding

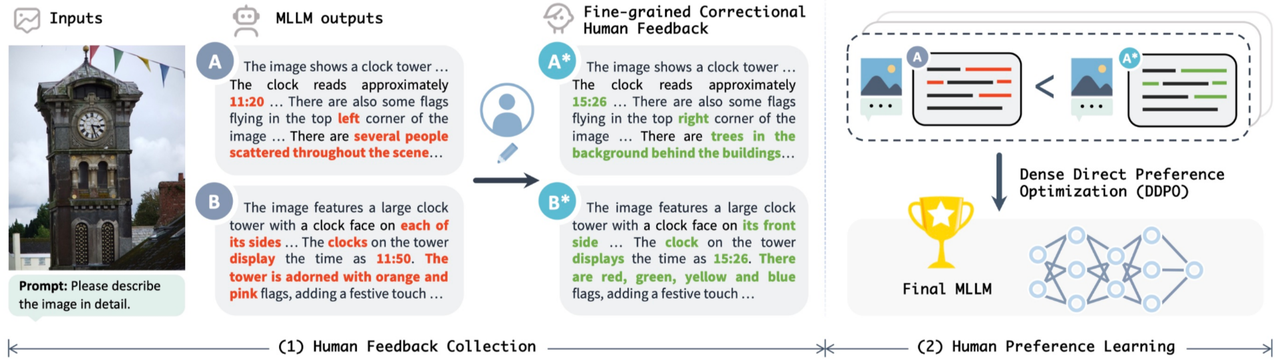


Dataset	Image Number	Paper Number	Image Category	Domain	Real Data
FigCAP (Chen et al., 2020)	219K	N / A	Bar, Line and Pie Charts	N / A	✗
SciCap (Yang et al., 2023b)	2.1M	295K	Open-Category	Computer Science and Machine Learning	✓
M-Paper (Hu et al., 2023)	350K	48K	Open-Category	Mainly "Deep Learning"	✓
ArXivCap (Ours)	6.4M	572K	Open-Category	Open-Domain	✓
FigureQA (Kahou et al., 2017)	140K	N / A	Bar, Line and Pie Charts	N / A	✗
DVQA (Kafle et al., 2018)	300K	N / A	Bar Charts	N / A	✗
ArXivQA (Ours)	32K	16.6K	Open-Category	Open-Domain	✓

Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models

Feedback Dataset for LLMs

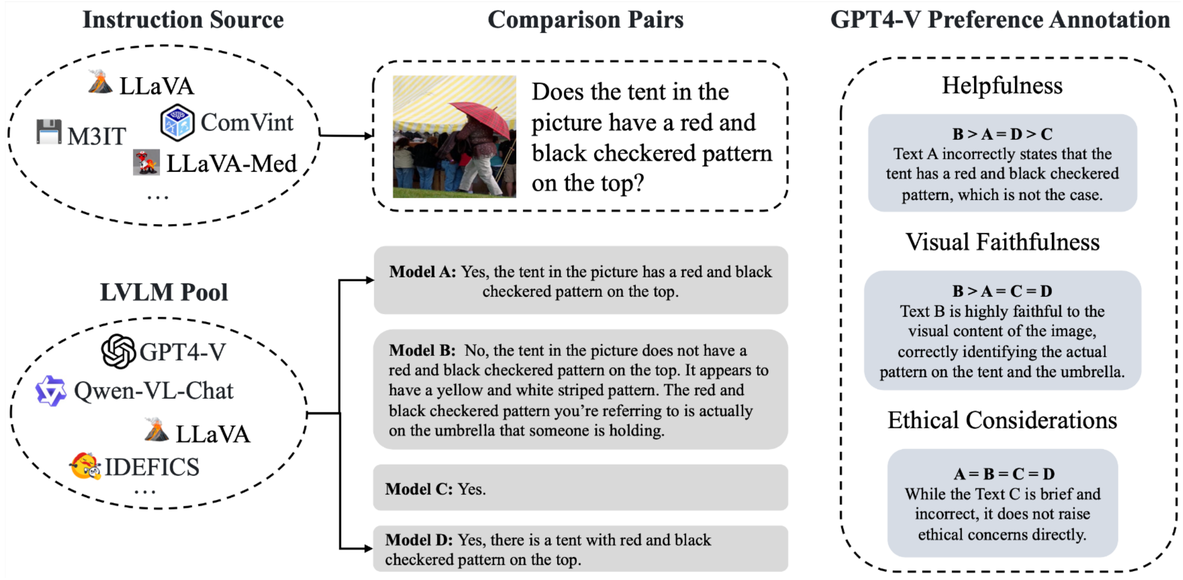




Aligning Large Multimodal Models with Factually Augmented RLHF

RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback

VLFeedback: Preference Distillation for LLMs



- We build a GPT-4V annotated vision-language feedback dataset, consists of 80K instructions decoded by various models, regarding three aspects.

A Guess of GPT-4V Recipe

- ChatGPT/GPT-4 as backbone LLMs for instruction following & reasoning
- Close-sourced Vision Encoders with 512px with tiling mechanism
  - Note CLIP is also from OpenAI
- Billion Scale (?) ChatGPT paraphrased detailed image-caption pairs for alignment training (inferred from DALLE & SORA practice)
- Diverse vision-language tasks collected for SFT
- OCR pipeline integrated (?)
- Professional alignment team for eliminating bias