Tokenizers for Unified Model

Paper Reading by Zhiying Lu

2025.06.18

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
- 方法1
- 方法2
- 总结反思

作者介绍

VILA-U: A Unified Foundation Model Integrating Visual Understanding and Generation

Yecheng Wu[1,2*] Zhuoyang Zhang[2*†] Junyu Chen[1,2] Haotian Tang[2†]
Dacheng Li[4†] Yunhao Fang[5†] Ligeng Zhu[3] Enze Xie[3]
Hongxu Yin[3] Li Yi[1] Song Han[2,3] Yao Lu[3]
Tsinghua University[1] MIT[2] NVIDIA[3] UC Berkeley[4] UC San Diego[5]
https://hanlab.mit.edu/projects/vila-u

Yecheng Wu

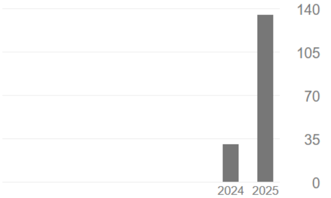
关注

Tsinghua University
在 mails.tsinghua.edu.cn 的电子邮件经过验证
Computer Vision

标题	引用次数	年份
Vila-u: a unified foundation model integrating visual understanding and generation Y Wu, Z Zhang, J Chen, H Tang, D Li, Y Fang, L Zhu, E Xie, H Yin, L Yi, ... arXiv preprint arXiv:2409.04429	100	2024
Hart: Efficient visual generation with hybrid autoregressive transformer H Tang, Y Wu, S Yang, E Xie, J Chen, J Chen, Z Zhang, H Cai, Y Lu, ... arXiv preprint arXiv:2410.10812	44	2024
Cot-vla: Visual chain-of-thought reasoning for vision-language-action models Q Zhao, Y Lu, MJ Kim, Z Fu, Z Zhang, Y Wu, Z Li, Q Ma, S Han, C Finn, ... Proceedings of the Computer Vision and Pattern Recognition Conference, 1702-1713	22	2025

引用次数

	总计	2020 年至今
引用	166	166
h 指数	3	3
i10 指数	3	3



Enze Xie

关注

NVIDIA Research, MMLab@HKU
在 connect.hku.hk 的电子邮件经过验证 - 首页
computer vision generative AI

引用次数

	总计	2020 年至今
引用	25709	25662
h 指数	45	45
i10 指数	74	74

标题
SegFormer: Simple and Efficient Design for Semantic Segmentation with E Xie, W Wang, Z Yu, A Anandkumar, JM Alvarez, P Luo Conference on Neural Information Processing Systems (NeurIPS), 2021
Pyramid vision transformer: A versatile backbone for dense prediction wit W Wang, E Xie, X Li, DP Fan, K Song, D Liang, T Lu, P Luo, L Shao IEEE International Conference on Computer Vision (ICCV)
PVT v2: Improved baselines with Pyramid Vision Transformer W Wang, E Xie, X Li, DP Fan, K Song, D Liang, T Lu, P Luo, L Shao Computational Visual Media 8 (3), 415-424

Song Han

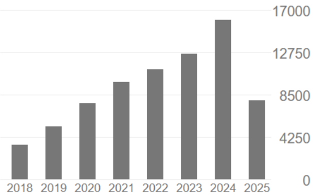
关注

Massachusetts Institute of Technology
在 mit.edu 的电子邮件经过验证 - 首页
Computer Architecture Deep Learning Computer Vision

标题	引用次数	年份
Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding S Han, H Mao, WJ Dally International Conference on Learning Representations (ICLR'16 best paper award)	12111	2015
SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5MB model size FN Iandola, S Han, MW Moskewicz, K Ashraf, WJ Dally, K Keutzer arXiv preprint arXiv:1602.07360	11337	2016

引用次数 查看全部

	总计	2020 年至今
引用	76391	65460
h 指数	74	72
i10 指数	148	145

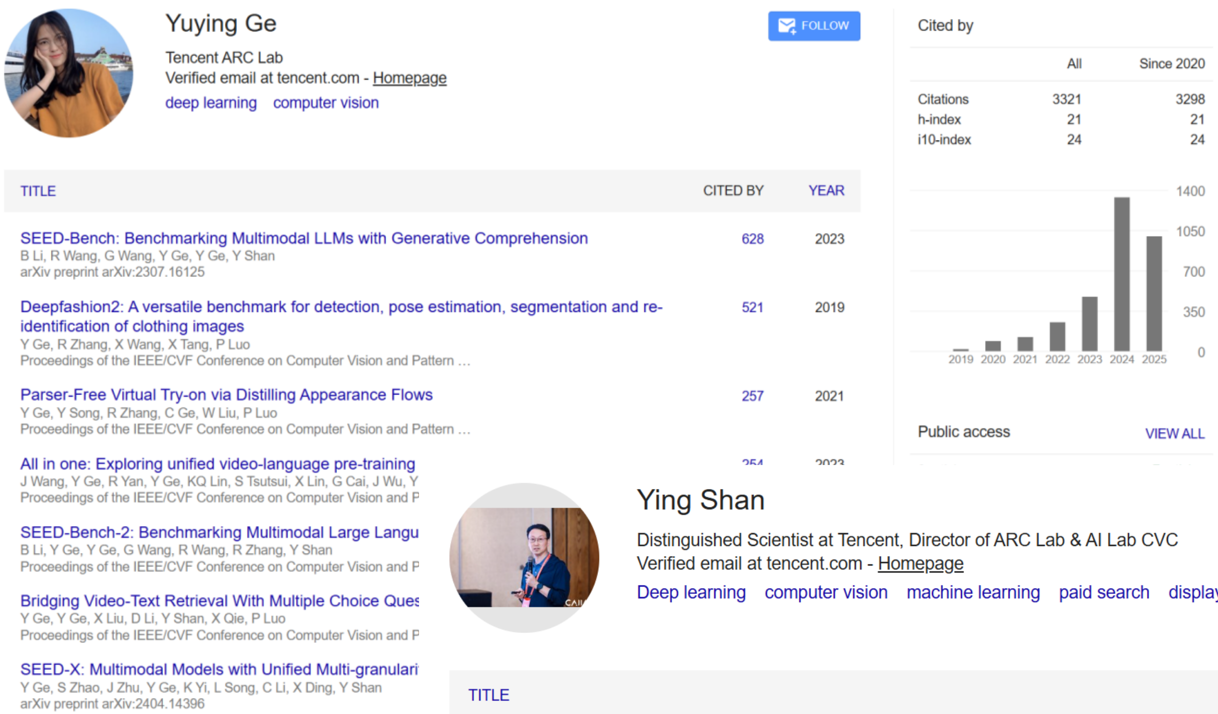


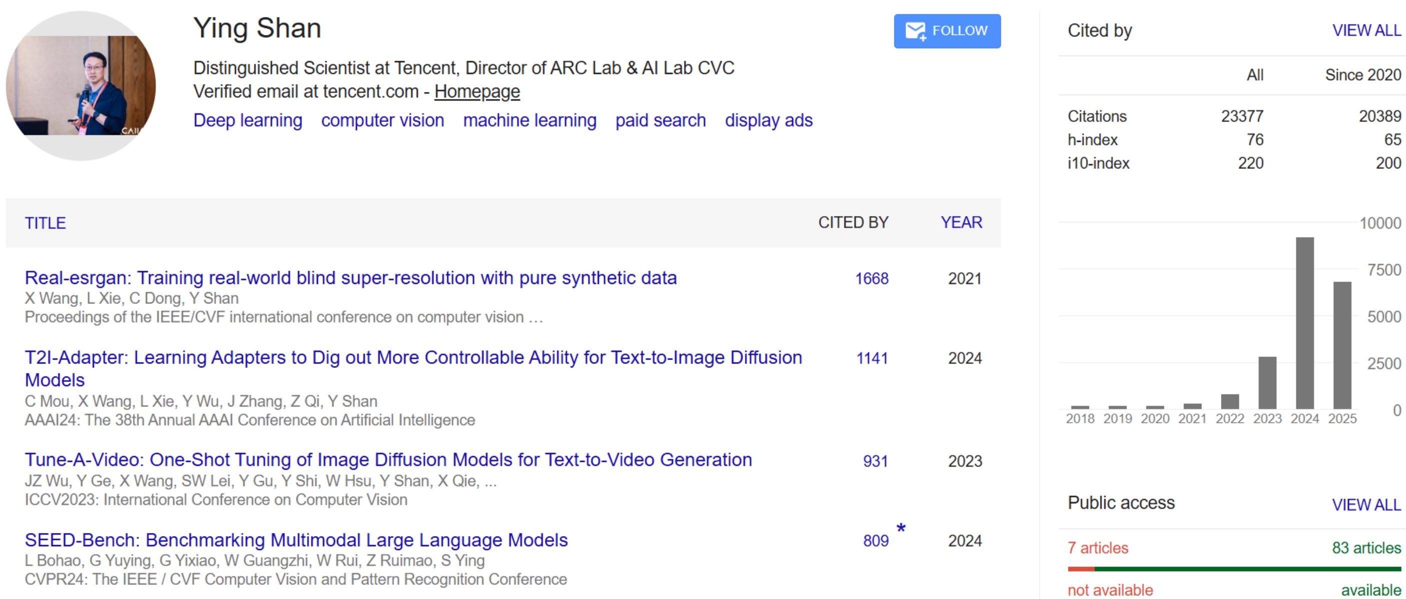
作者介绍

Divot: Diffusion Powers Video Tokenizer for Comprehension and Generation

Yuying Ge Yizhuo Li Yixiao Ge Ying Shan
ARC Lab, Tencent PCG

https://github.com/TencentARC/Divot



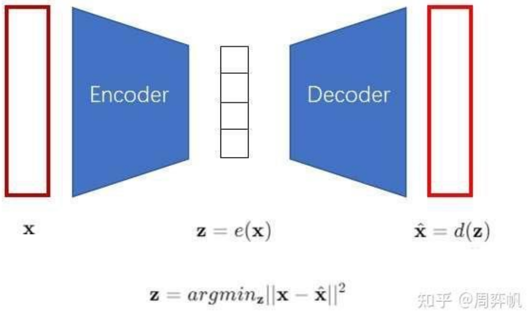


- 作者介绍
- 背景介绍
- 方法1
- 方法2
- 总结反思

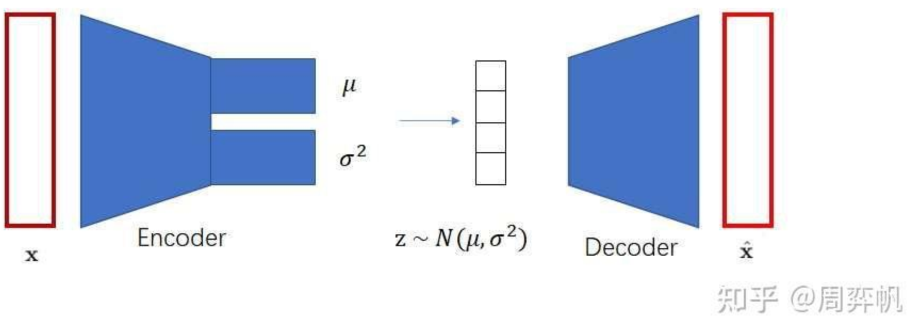
背景知识

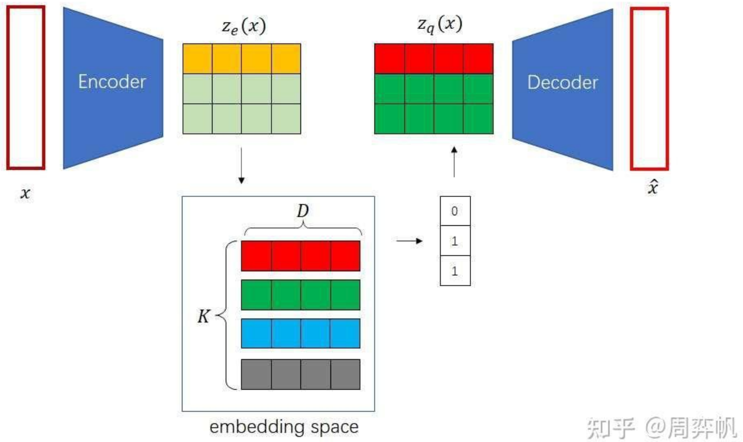
- Tokenizer是一种广义的概念，也是所有任务的最上游
- Tokenizer主要将原始输入映射到某种特定的空间，这个空间可以称为latent space（隐空间）或者codebook（码本）
- 基于这个特定的空间，后续模型可以实现各种下游任务

背景知识



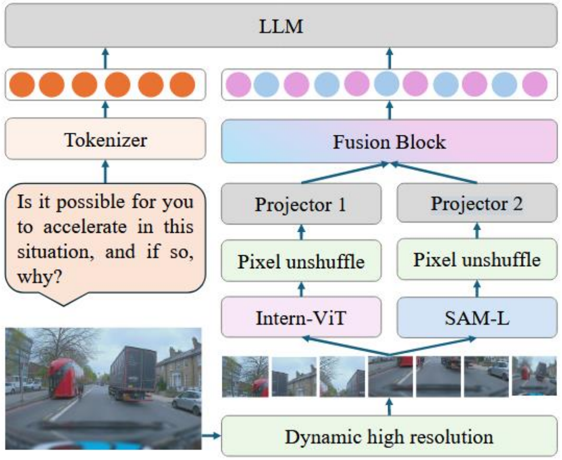
- 在NLP中主要为码本的概念，即每个字符或者字母组合都有自己的嵌入嵌入向量，常见的有BPE编码器等
- 在视觉领域最常见的是VAE，将视觉信息压缩为连续高斯分布
- 在多模态领域，CLIP等也可以被认为是一种Tokenizer，常用于各种多模态下游任务和MLLM等
- 实际上，任何foundation model都可以被认为是一种Tokenizer！

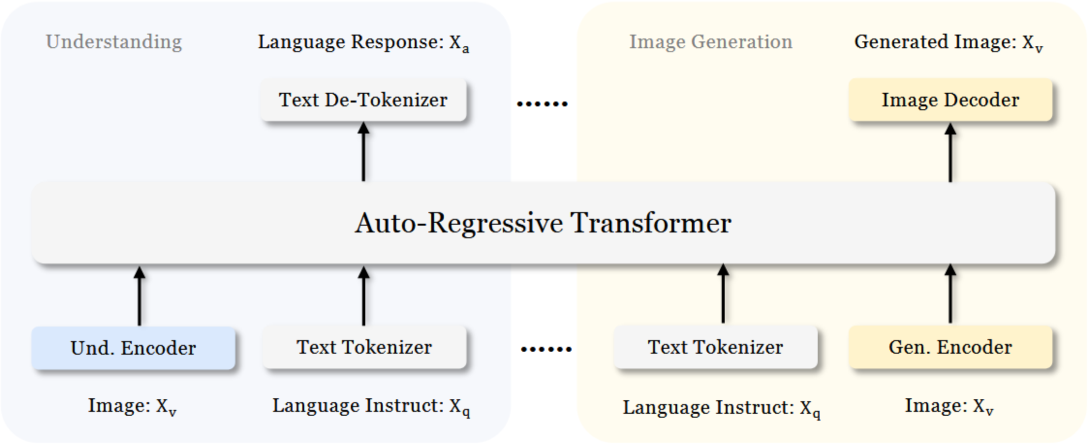




背景知识

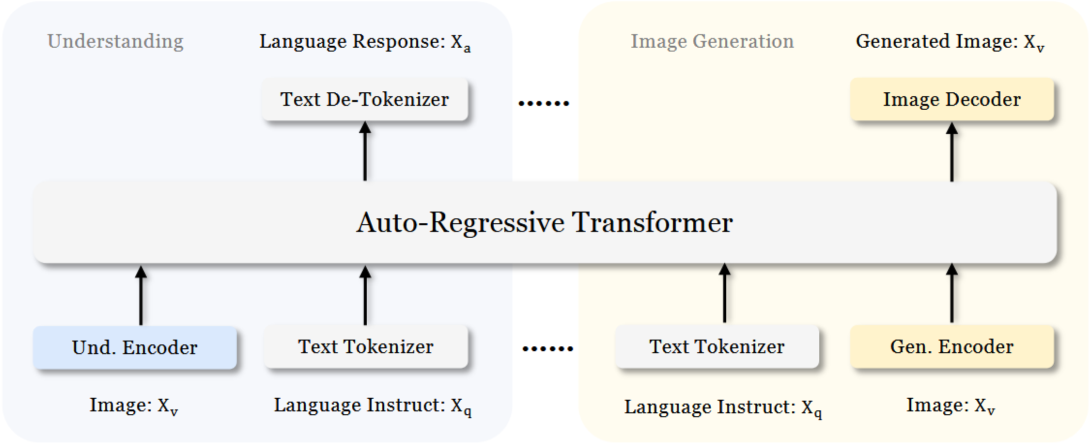
- Tokenizer在更多情况下是为了特定任务和特定模态实现的，因此很多时候增加了非常多的先验信息
- 在2023-2024年的MLLM领域，众多工作都在考虑如何mixture of vision expert来赋予各种视觉理解的先验
- 来到统一模型时代，如何同时均衡视觉生成和视觉理解两种任务，是设计Tokenizer的关键





背景知识

- 满足理解任务，采用CLIP，SigLIP（2）即可
- 满足生成任务，则需要采用VAE/VQ-VAE等，因为需要对输出进行解码
- Janus中首先提出了**解耦**的概念，符合利用现有各种Tokenizer的先验进行编码的思路，并且能做到两种模型互不干扰
- 是否能设计出单一编码器，同时实现两种功能?
- 涉及到两个关键概念：连续vs离散，低级vs高级

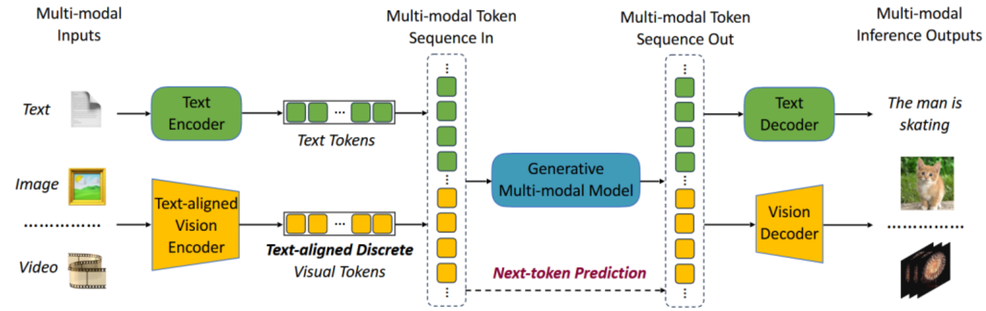


背景知识

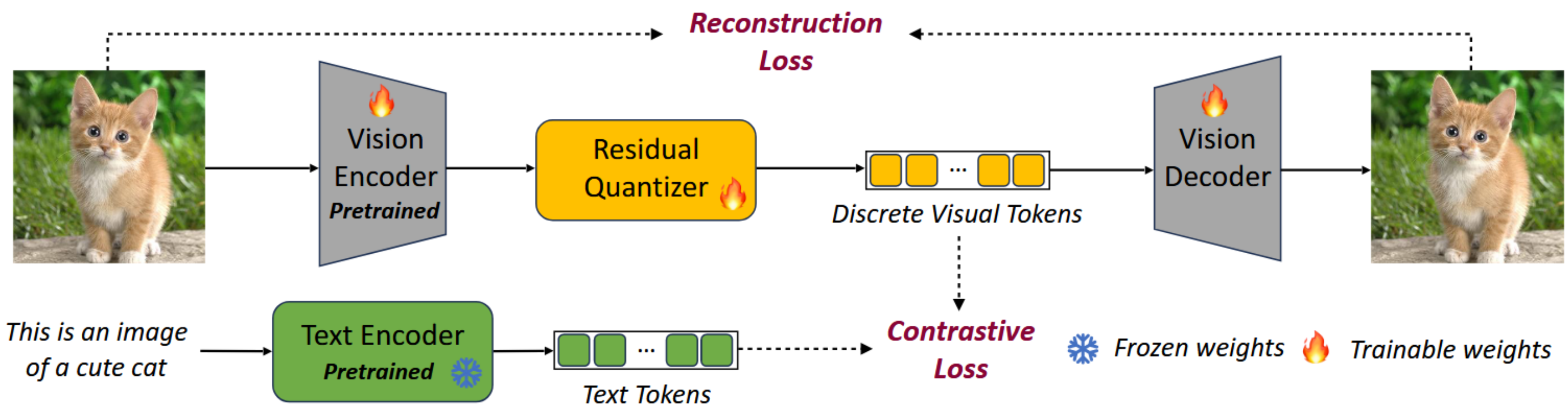
- 由于LLM的输出预测是离散的，因此采用连续码本很难做到预测，这使得大部分人抛弃了原本VAE，转而采用VQ或者是1d tokenizer
- 由于像素级信息比较低级，而LLM是在高级语义信息上进行处理，因此大部分人考虑将像素级信息隐含到tokenizer中，而输出只有高级语义
- 进而诞生出两条路线
  - CLIP编码器+VQ-VAE
  - 1d tokenizer+Diffusion Decoder

- 作者介绍
- 背景介绍
- 方法1
- 方法2
- 总结反思

VILA-U



- 单一tokenizer设计，采用离散化的生成码本，加载CLIP的双编码器
- 采用RQ-VAE设计进行量化，使得表征层级更丰富
- RQ-VAE具有残差量化性，重复量化同一特征，每次减去之前的量化值 $\hat{\mathbf{z}} = \sum_{i=1}^{D} \mathbf{e}(k_i).$



$$\mathcal{RQ}(\mathbf{z};\mathcal{C},D) = (k_1,\cdots,k_D)\in[K]^D, \qquad \mathcal{Q}(\mathbf{z};\mathcal{C}) = \underset{k\in[K]}{\arg\min}\|\mathbf{z}-\mathbf{e}(k)\|_2^2.$$

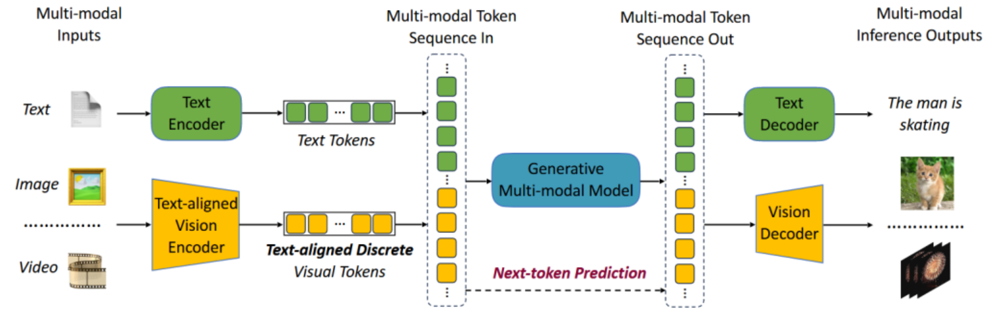
$$k_d = \mathcal{Q}\left(\mathbf{r}_{d-1},\mathcal{C}\right),$$

$$\mathbf{r}_d = \mathbf{r}_{d-1} - \mathbf{e}\left(k_d\right),$$

$$\mathcal{L}_{total} = w_{contra}\mathcal{L}_{contra} + w_{recon}\mathcal{L}_{recon}$$

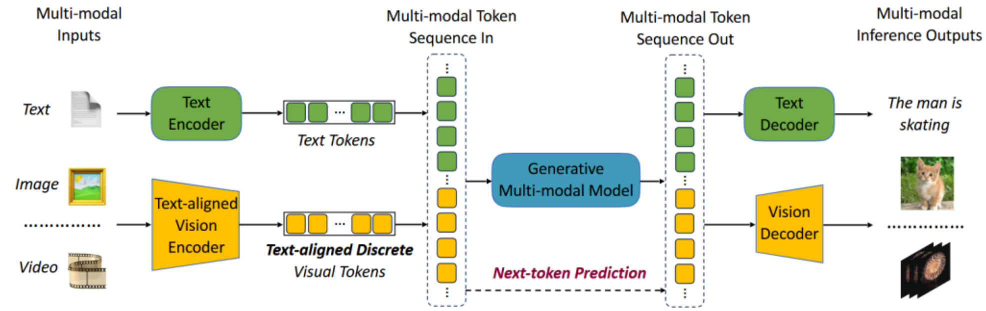
$$\mathcal{L}_{\text{text}} = -\sum_{i=1}^{T}\log P_\theta\left(y_i|y_{<i}\right), \quad \mathcal{L}_{\text{visual}} = -\sum_{j=1}^{T}\sum_{d=1}^{D}\log P_\delta\left(k_{jd}|k_{j,<d}\right),$$

VILA-U



- 四种训练失败的策略：
  - （1）只加载clip文本编码器权重—从头训练CLIP过于困难
  - （2）加载RQ-VAE权重到编解码器，但其他从头训练—同上，且batchsize难平衡
  - （3）固定视觉编码器—难以学习到低级视觉信息
  - （4）不固定文本编码器—量化过程破坏了语义对齐性

VILA-U

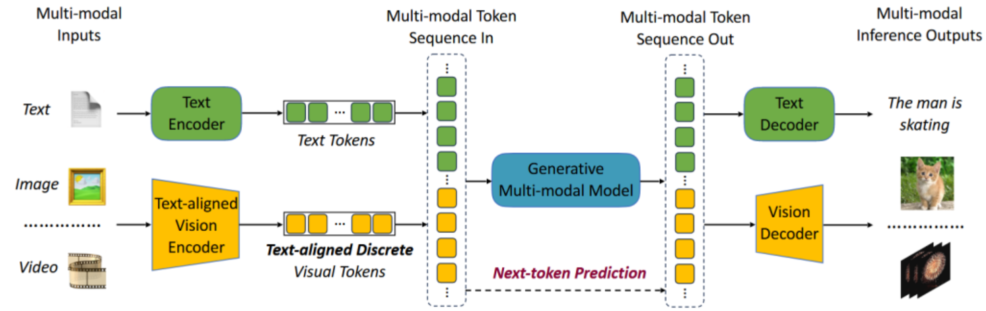


Model	Pretrained Weights	Resolution	Shape of Code	rFID↓	Top-1 Accuracy↑
VQ-GAN [22]	–	256 × 256	16 × 16	4.98	–
RQ-VAE [33]	–	256 × 256	8 × 8 × 4	3.20	–
RQ-VAE [33]	–	256 × 256	16 × 16 × 4	1.30	–
Ours	SigLIP-Large	256 × 256	16 × 16 × 4	1.80	73.3
Ours	SigLIP-SO400M	384 × 384	27 × 27 × 16	1.25	78.0

Method	LLM	Visual Token	Res.	VQAv2	GQA	TextVQA	POPE	MME	SEED	MM-Vet
LLaVA-1.5 [51]	Vicuna-1.5-7B	Continuous	336	78.5*	62.0*	58.2	85.9	1510.7	58.6	30.5
VILA [45]	LLaMA-2-7B	Continuous	336	79.9*	62.3*	64.4	85.5	1533.0	61.1	34.9
Unified-IO 2 [52]	6.8B from scratch	Continuous	384	79.4*	–	–	87.7	–	61.8	–
InstructBLIP [15]	Vicuna-7B	Continuous	224	–	49.2	50.1	–	–	53.4	26.2
IDEFICS-9B [32]	LLaMA-7B	Continuous	224	50.9	38.4	25.9	–	–	–	–
Emu [64]	LLaMA-13B	Continuous	224	52.0	–	–	–	–	–	–
LaVIT [31]	LLaMA-7B	Continuous	224	66.0	46.8	–	–	–	–	–
DreamLLM [19]	Vicuna-7B	Continuous	224	72.9*	–	41.8	–	–	–	36.6
Video-LaVIT [30]	LLaMA-2-7B	Continuous	224	80.2*	63.6*	–	–	1581.5	64.4	35.0
CM3Leon-7B [75]	7B from scratch	Discrete	256	47.6	–	–	–	–	–	–
LWM [48]	LLaMA-2-7B	Discrete	256	55.8	44.8	18.8	75.2	–	–	9.6
Show-o [70]	Phi-1.5-1.3B	Discrete	256	59.3*	48.7*	–	73.8	948.4	–	–
Ours	LLaMA-2-7B	Discrete	256	75.3*	58.3*	48.3	83.9	1336.2	56.3	27.7
Ours	LLaMA-2-7B	Discrete	384	79.4*	60.8*	60.8	85.8	1401.8	59.0	33.5

Method	LLM	Visual Token	Res.	MSVD-QA	MSRVTT-QA	TGIF-QA	Activity Net-QA
Unified-IO 2 [52]	6.8B from scratch	Continuous	384	52.1	42.5	–	–
Emu [64]	LLaMA-13B	Continuous	224	–	18.8	8.3	–
VideoChat [40]	Vicuna-7B	Continuous	224	56.3	45	34.4	–
Video-LLaMA [78]	LLaMA-2-7B	Continuous	224	51.6	29.6	–	–
Video-ChatGPT [53]	LLaMA-2-7B	Continuous	224	64.9	49.3	51.4	35.2
Video-LLava [44]	Vicuna-7B	Continuous	224	70.7	59.2	70.0	45.3
Video-LaVIT [30]	LLaMA-2-7B	Continuous	224	73.5	59.5	–	50.2
LWM [48]	LLaMA-2-7B	Discrete	256	55.9	44.1	40.9	–
Ours	LLaMA-2-7B	Discrete	256	73.4	58.9	51.3	51.6
Ours	LLaMA-2-7B	Discrete	384	75.3	60.0	51.9	52.7

VILA-U



Method	Type	#Training Images	Attribute↑	Scene↑	Relation↑			Overall↑
					Spatial	Action	Part	
SD v2.1 [60]	Diffusion	2000M	0.80	0.79	0.76	0.77	0.80	0.78
SD-XL [57]	Diffusion	2000M	0.84	0.84	0.82	0.83	0.89	0.83
Midjourney v6 [59]	Diffusion	–	0.88	0.87	0.87	0.87	0.91	0.87
DALL-E 3 [47]	Diffusion	–	0.91	0.90	0.92	0.89	0.91	0.90
Show-o [70]	Discrete Diff.	36M	0.72	0.72	0.70	0.70	0.75	0.70
LWM [48]	Autoregressive	–	0.63	0.62	0.65	0.63	0.70	0.63
Ours (256)	Autoregressive	15M	0.78	0.78	0.77	0.78	0.79	0.76
Ours (384)	Autoregressive	15M	0.75	0.76	0.75	0.73	0.75	0.73

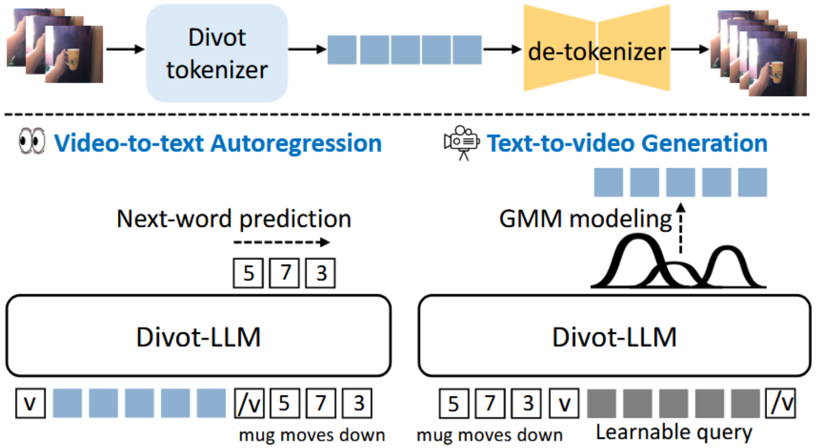
Pretrained Weights	Data size	Loss Type	Top-1 Accuracy	VQAv2	POPE	MME	SEED	MM-Vet
SigLIP-Large	25M	Recon.	–	57.7	75.1	937.7	38.7	15.3
SigLIP-Large	25M	Recon. + Contra.	62.9	68.0	83.7	1219	50.4	20.8
SigLIP-Large	700M	Recon. + Contra.	73.3	75.3	83.9	1336.2	56.3	27.7

Table 7: Impact of contrastive loss to visual generation.

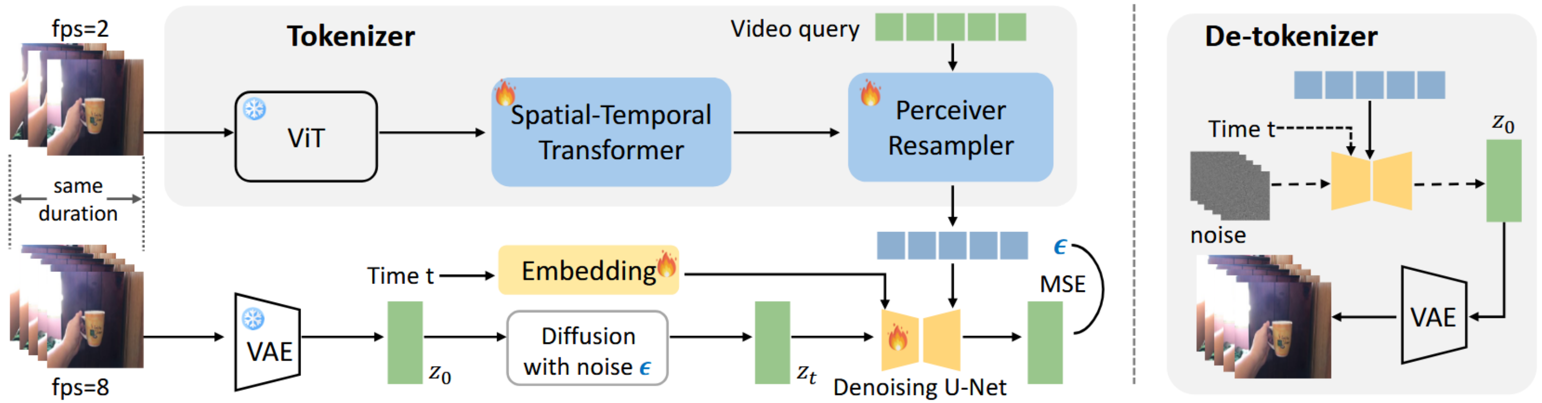
Vision Tower	LLM	Resolution	rFID ↓	FID ↓
RQ-VAE [33]	Sheared-LLaMA-1.3B	256 × 256	1.30	12.0
Ours	Sheared-LLaMA-1.3B	256 × 256	1.80	13.2

- 作者介绍
- 背景介绍
- 方法1
- 方法2
- 总结反思

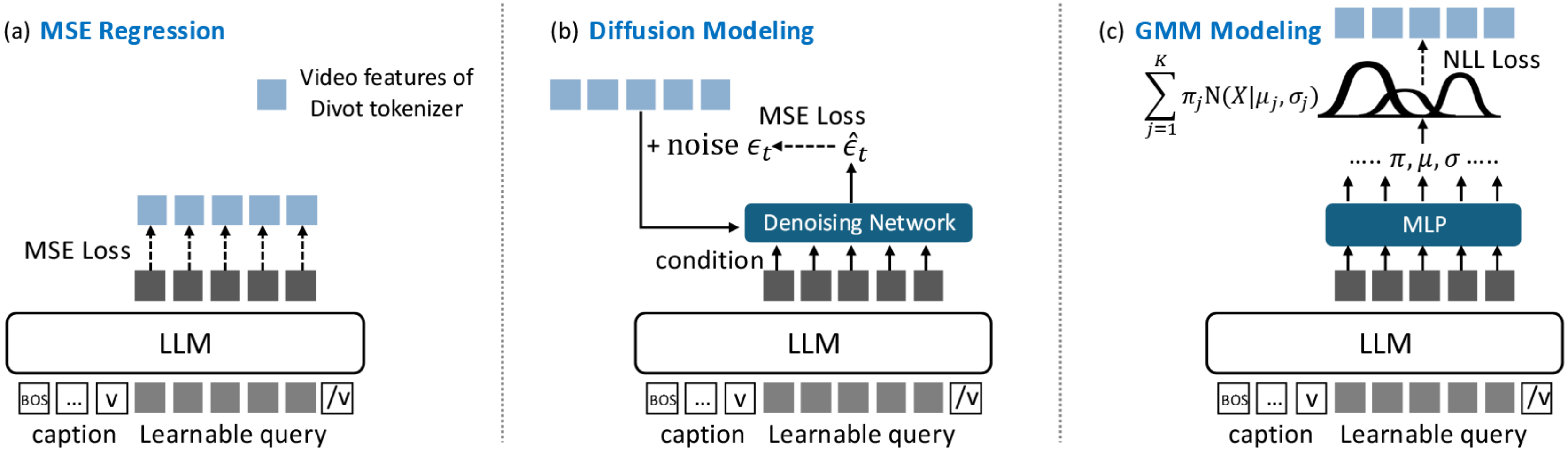
Divot



- 1d-tokenizer设计+diffusion decoder
- 没有量化过程，因此LLM输出采用混合高斯
- 训练编码器时引入diffusion监督
- **主要思想：对于同一视频，如果tokenizer提供的特征作为条件能够使得diffusion预测出噪声，则认为该tokenizer成功捕捉了时间和空间信息**
- 采用diffusion方式进行**视觉自监督**地训练一个tokenizer



Divot



- 训练LLM时候关于生成任务的监督
- 如果对high-level video特征引入噪声效果会变差，说明高级语义特征需要精确建模
- 在视觉特征中采用双向建模更合适

	Representation		Objective				Mechanism		
	patch-position dependent	patch-position independent	MSE	Diffusion		GMM	AR	Query	
				ϵ-pred	v-pred			causal	bidirectional
CLIPSIM (↑)	0.3192	**0.3265**	0.3168	0.2811	0.2842	**0.3265**	0.2386	0.3080	**0.3265**
FVD (↓)	378.50	**366.60**	438.94	418.19	377.17	**366.60**	447.88	416.60	**366.60**

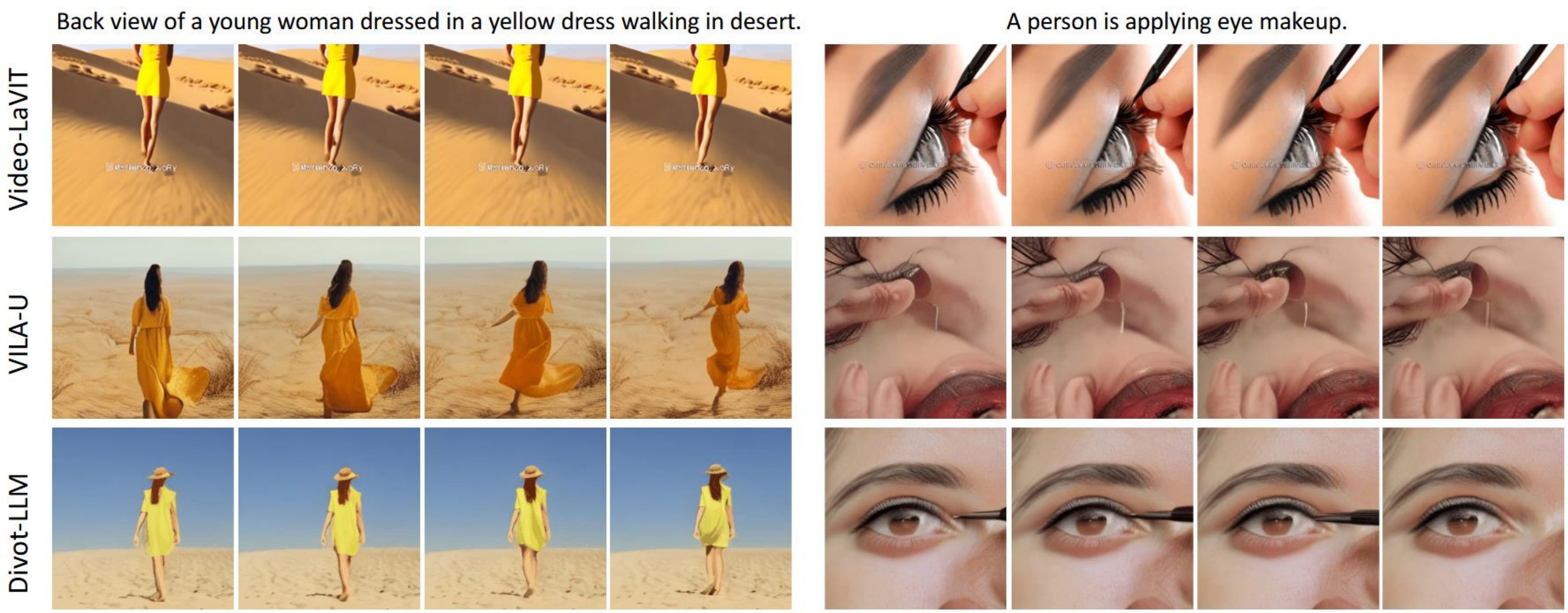
Divot

Stage	Type	Dataset
Tokenize	Pure Video	WebVid-10M [2], Panda-70M [9]
Pre-train	Video-text	WebVid-10M [2]
	Image-text	CC3M [52], CapsFusion [87], LAION-COCO [51]
SFT	Classification	Kinetics-710 [27], SSV2 [18]
	VQA	TGIF [34], NextQA [76], CLEVRER [85], YouCook2 [92], PerceptionTest[48], EgoQA [19], ActivityNetQA[88]
	Instruction	Video-ChatGPT[43], LLaVA-mixed[39], Valley [42], LLaVA-Video-178K[37]
	Generation	WebVid-10M [2]
	StoryTelling	In-house data

Model	Data size	Unified	MSR-VTT	
			CLIPSIM (↑)	FVD (↓)
CogVideo [21]	5.4M	×	0.2631	1294
Video LDM [5]	10M	×	0.2929	-
VideoComposer [66]	10M	×	0.2932	580
InternVid [68]	28M	×	0.2951	-
Make-A-Video [53]	20M	×	**0.3049**	-
VideoPoet [29]	270M	×	**0.3049**	213
PYoCo [14]	22.5M	×	-	-
SVD [4]	152M	×	-	-
Video-LavIT [26]	10M	✓	<u>0.3012</u>	<u>188.36</u>
Loong [69]	16M	×	0.2903	274
Snap Video [45]	-	×	0.2793	**110.4**
VILA-U [74]	1M	✓	0.2937	499.06
Divot-LLM	4.8M	✓	0.2938	301.4

Model	LLM size	Video-Gen	EgoSchema	Perception-Test	MVBench	MSVD	ActivityNet
Gemini 1.0 Pro [58]	-	×	55.7	51.1	-	-	49.8
Gemini 1.5 Pro [59]	-	×	63.2	-	-	-	56.7
GPT4-V [46]	-	×	55.6	-	43.7	-	59.5
GPT4-O [47]	-	×	**72.2**	-	-	-	<u>61.9</u>
LLaMA-VID [35]	7B	×	38.5	44.6	41.9	69.7	47.4
Video-ChatGPT [43]	7B	×	-	-	-	64.9	35.2
Video-LLaVA [37]	7B	×	38.4	44.3	41.0	70.7	45.3
VideoChat2 [31]	7B	×	42.2	47.3	51.1	70.0	49.1
LLaVA-NeXT-Video [38]	7B	×	43.9	48.8	46.5	67.8	53.5
LLaVA-NeXT-Video [38]	32B	×	60.9	-	-	-	54.3
PLLaVA [81]	34B	×	-	58.1	-	-	60.9
LLaVA-OneVision [30]	72B	×	62.0	-	-	-	**62.3**
VideoLLaMA2 [10]	7B	×	51.7	51.4	<u>54.6</u>	70.9	50.2
VideoLLaMA2 [10]	72B	×	<u>63.9</u>	<u>57.5</u>	**62.0**	71.0	55.2
LWM [40]	7B	✓	-	-	-	55.9	-
Video-LaVIT [26]	7B	✓	37.3	47.9	-	73.2	50.1
VILA-U [74]	7B	✓	-	-	-	<u>75.3</u>	52.7
Divot-LLM	7B	✓	46.5	**58.3**	52.1	**76.4**	55.8

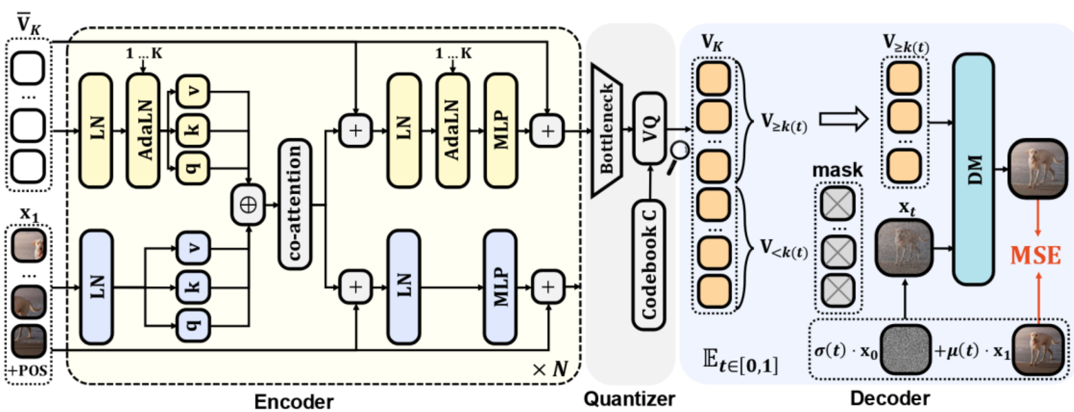
Divot

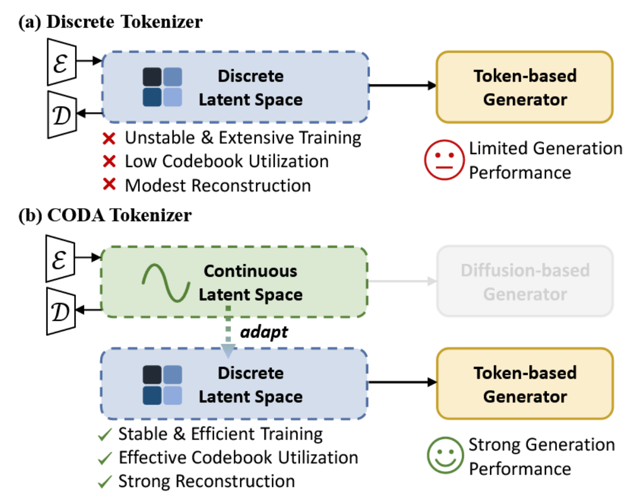

Back view of a young woman dressed in a yellow dress walking in desert.
A person is applying eye makeup.
Video-LaVIT
VILA-U
Divot-LLM

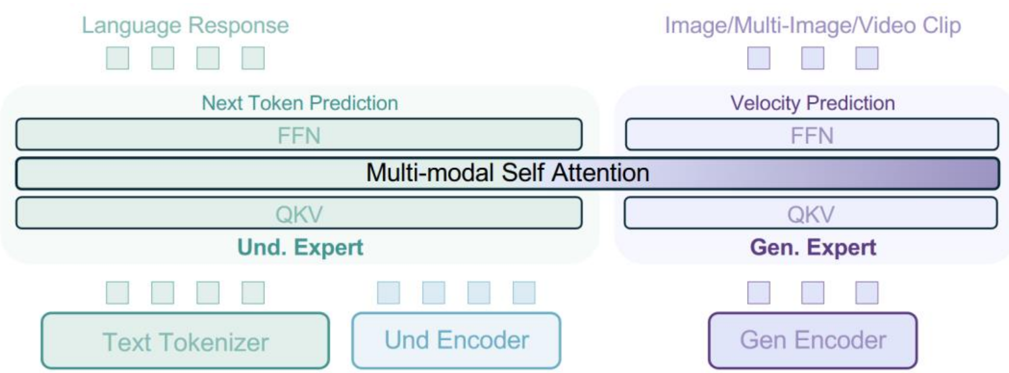
- 作者介绍
- 背景介绍
- 方法1
- 方法2
- 总结反思

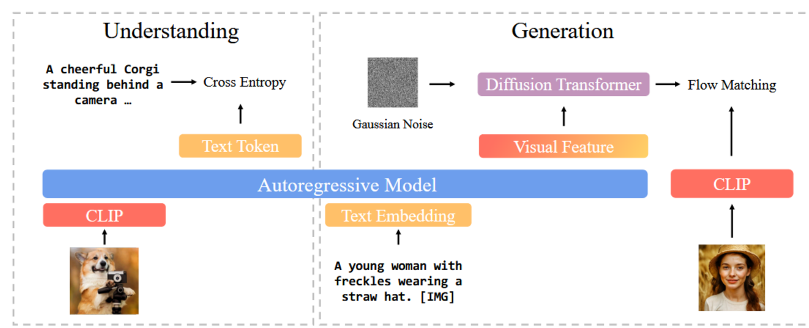
总结反思

- 1d tokenizer是统一编码器的主流实现，解耦编码器也仍在广泛实践
- 如何同时构建具有低级和高级语义的空间是关键









谢谢!