专题综述

多模态生成式大模型的自回归式建模

金阳 孙至诚 穆亚东
北京大学

本文主要介绍北京大学与快手科技公司合作研究的成果，发表在ICLR 2024的工作LaVIT[1]和ICML 2024的工作Video-LaVIT[2]，代表了基于自回归架构的多模态生成式大模型的最新研究进展。

一、研究背景

生成式人工智能是近年来学术界和工业界的研究热点。美国人工智能公司 OpenAI 在 2024 年初发布文生视频模型 Sora。该模型只需要读入一段提示文本（即所谓的“prompt”），即能自动生成包含提示所描述的复杂场景背景、多角色交互、动作语义的高清视频。在不少示例中，生成视频的质量会随着提示文本中的描述性细节的丰富而得到提升。与现有的文生视频模型如 Runway Gen-2、Pika 等相比较，Sora 在多个样例视频中展示了长达一分钟的连续、稳定、高品质视频，而现有其他模型通常仅能产生几秒钟连贯性的视频输出。Sora 系统的发布，普遍被认为是继 ChatGPT 之后生成式人工智能发展上的又一个重要技术突破。其背后的核心技术，包括视频时空表征学习和基于 Transformer 的扩散模型等，不仅在文生视频这一任务中取得当前最佳性能，也可以被预见将被迁移至更多模态的复合型内容生成任务中，将迅速地改变当前广告、互动娱乐、影视制作和媒体宣传等行业的技术生态，成为推动社会经济发展的重要技术之一。

近年来扩散模型和自回归模型逐渐成为相关任务的主流技术。针对文生图任务，OpenAI 公司在 2021 年发布了基于 VQ-VAE 和 GPT 的 DALL-E[3]，之后又发布了改进版本 DALL-E2[4]和 DALL-E3[5]，国内的清华大学等单位提出了 CogView[6]模型，通过对文本和图像同时进行大规模的协同预训练来实现更精确的文生图。隐扩散模型（latent diffusion）[7]和 GLIDE[8]提出了基于文本指导的图像扩散过程，经验性对比了 CLIP 指导

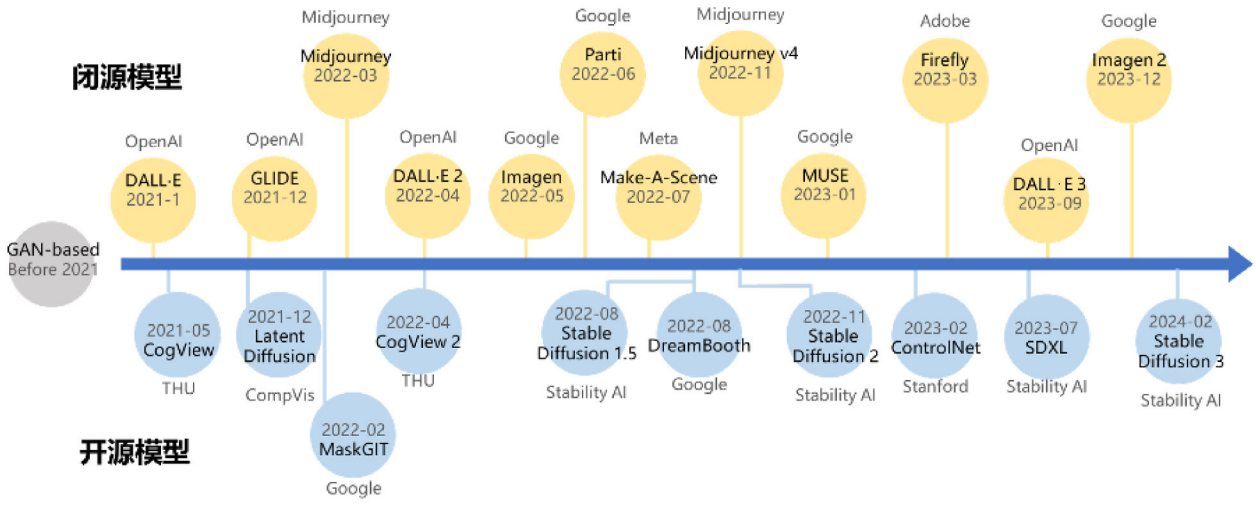


图 1　文本生成图像模型的技术发展时间线

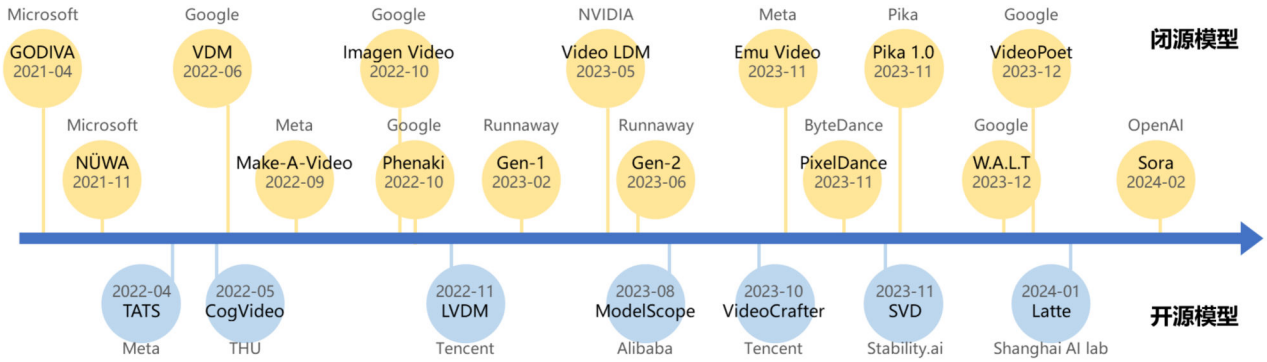


图 2　视频生成式模型的技术发展时间线

和无分类器指导（classifier-free guidance）两种策略的效果，验证了后者在图片真实性和主题相似方面效果更好。其他模型还包括 MaskGIT[9]，改进模型 CogView2[10]，Stable Diffusion 及其改进版本，Imagen[11]及其改进版本 Imagen2，闭源模型 Parti[12]、Make-A-Scene[13]、DreamBooth[14]、MUSE[15]、Midjourney、Firefly 等。ControlNet[16]通过添加额外控制条件，来引导 Stable Diffusion 按照创作者的创作思路生成图像，从而提升图像生成的可控性和精度。该方向的技术发展回顾见图 1。

针对文生视频任务，主流方法主要基于扩散模型，如 LVDM[17]、Stable Video Diffusion[18]等。视频生成式模型可以选择不同的提示信息作为输入，如文本描述或单幅图像。其中，[18]旨在利用静态图像作为条件帧，从而实现基于此单一图像输入的视频生成。在 2021 年提出的多任务模型 “女娲” NÜWA[19]以文本或视觉草图作为输入的自适应编码器和由 8 个视觉合成任务共享的解码器组成，可以支持涂鸦生成图像、图像填充、涂鸦转视频等 8 种模式。为了更好地支持多任务的统一建模，自回归模型（如 EMU-Video[20]、VideoPoet[21]等）自 2023 年也逐渐受到关注，这些方法首先对不同模态各自进行离散标记化，继而送入 LLM 风格的自回归模型进行多任务的统一训练。图 2 展示了相关技术近来的发展时间线。

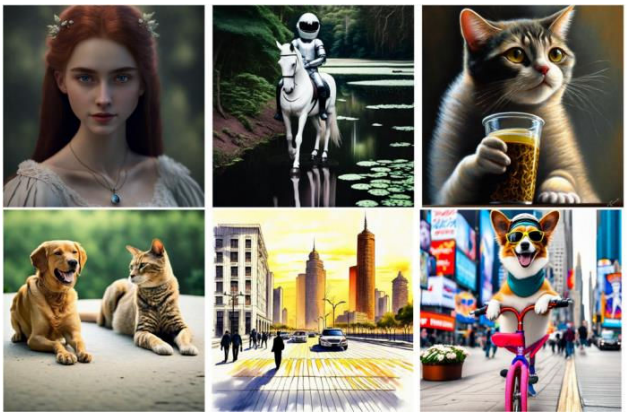


图 3　LaVIT 模型的文生图结果示例

图 4　LaVIT 模型的多模态生成示例

二、自回归文生图大模型LaVIT

1．模型简介

LaVIT 作为一个新型的通用多模态基础模型，可以

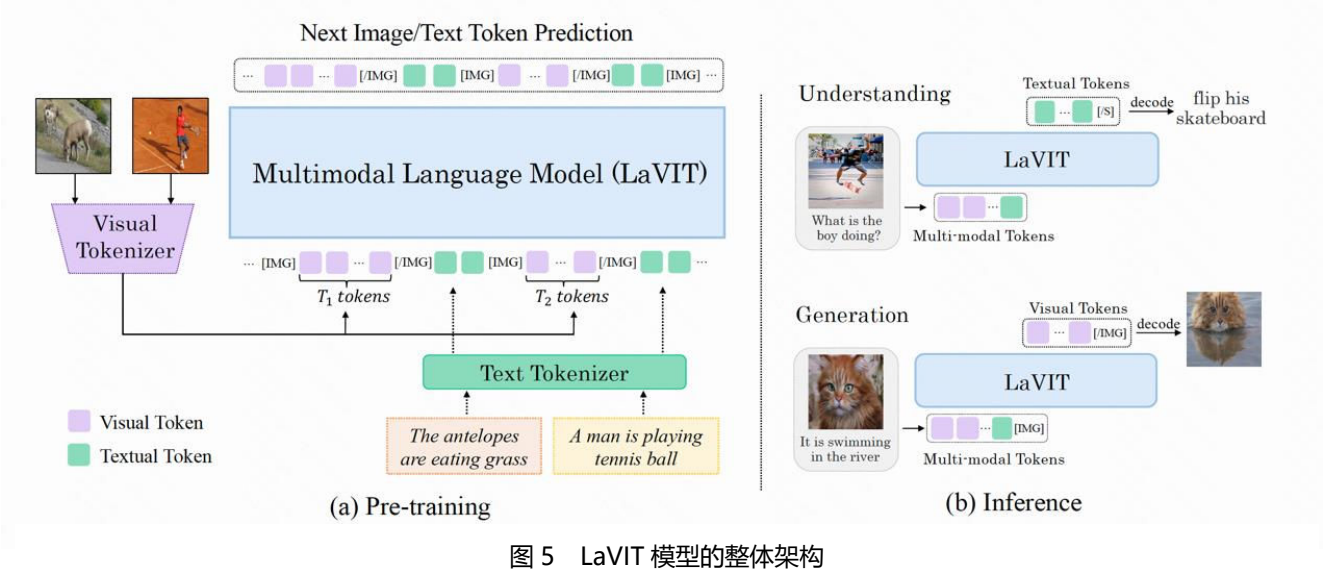


图 5　LaVIT 模型的整体架构

像语言那样，既能够理解也能生成视觉内容。LaVIT 继承了大语言模型成功的训练范式，即以自回归的方式预测下一个图像或文本 token。在训练完成后，其可以充当一个多模态通用接口，无需进一步的微调，就能执行多模态理解和生成任务。例如，LaVIT 具有以下的能力：

（1）实现高质量文本到图像的生成：LaVIT 能够根据给定的文本提示生成高质量、多种纵横比和高美感的图像（如图 3）。其图像生成能力与最先进的图像生成模型（如 Parti、SDXL 和 DALL-E3）相媲美；

（2）根据多模态提示进行图像生成：由于在 LaVIT 中，图像和文本都被统一表示为离散化的 token，因此其可以接受多种模态组合（例如图 4 中的文本、图像+文本、图像+图像）作为提示，生成相应的图像，而无需进行任何微调。

（3）理解图像内容并回答问题：在给定输入图像的情况下，LaVIT 能够阅读图像内容并理解其语义。例如，模型可以为输入的图像提供 caption 并回答相应的问题。

2．模型架构

LaVIT 模型的整体架构如图 5 所示，其优化过程包括两个阶段：

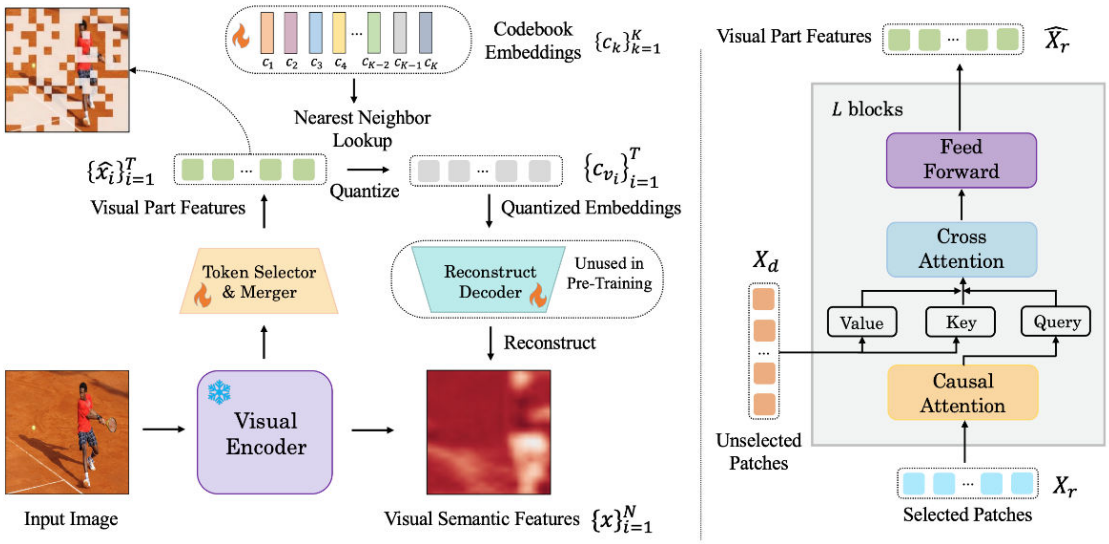
阶段 1：动态视觉分词器

为了能够像自然语言一样理解和生成视觉内容，LaVIT 引入了一个设计良好的视觉分词器，用于将视觉内容（连续信号）转换为像文本一样的 token 序列，就像 LLM 能够理解的外语一样。作者认为，为了实现统一视觉和语言的建模，该视觉分词器（Tokenizer）应该具有以下两个特性：

离散化：视觉 token 应该被表示为像文本一样的离散化形式。这样对于两种模态采用统一的表示形式，有利于 LaVIT 在一个统一的自回归生成式训练框架下，使用相同的分类损失进行多模态建模优化。

动态化：与文本 token 的不同的是，图像 patch 之间有着显著的相互依赖性，这使得从其他图像 patch 中推断另一个 patch 相对简单。因此，这种依赖性会降低原本 LLM 的 next-token prediction 优化目标的有效性。LaVIT 提出通过使用 token merging 来降低视觉 patch 之间的冗余性，其根据不同图像语义复杂度的不同，编码出动态的视觉 token 数量。这样对于复杂程度不同的图像，采用动态的 token 编码也进一步提高了预训练的效率，避免了冗余的 token 计算。

图 6 展示了 LaVIT 所提出的视觉分词器结构。该动态视觉分词器包括 token 选择器和 token 合并器。如图 6 所示，token 选择器用来选择最具信息的图像区块，而 token 合并器则将那些 uninformative 的视觉块的信息压缩到保留下的 token 上，实现对冗余 token 的 merging。整个动态视觉分词器则通过最大限度地重构输入图像的语义进行训练。



(a) Dynamic Visual Tokenizer　　(b) The Token Merger

图 6　LaVIT 模型的动态视觉 token 生成器（左）和 token 合并器（右）

Token 选择器：Token 选择器接收 N 个图像区块级的特征作为输入，其目标是评估每个图像区块的重要性并选择信息量最高的区块，以充分代表整个图像的语义。为实现这一目标，采用轻量级模块，由多个 MLP 层组成，用于预测分布π。通过从分布 π 中采样，生成一个二进制决策 mask，用于指示是否保留相应的图像区块。

Token 合并器：Token 合并器据生成的决策掩码，将 N 个图像区块划分为保留X_r和舍弃X_d两组。与直接丢弃X_d不同，token 合并器可以最大限度地保留输入图像的详细语义。token 合并器由 L 个堆叠的块组成，每个块包括因果自注意力层、交叉注意力层和前馈层。因果自注意力层中，X_r中的每个 token 只关注其前面的 token，以确保与 LLM 中的文本 token 形式一致。与双向自注意相比，这种策略表现更好。交叉注意力层将保留的 token X_r作为 query，并根据它们在语义上的相似性合并X_d中的 token。

阶段 2：统一的生成式预训练

经过视觉分词器处理后的视觉 token 与文本 token 相连接形成多模态序列作为训练时的输入。为了区分两种模态，作者在图像 token 序列的开头和结尾插入了特殊 token：[IMG] 和 [/IMG]，用于表示视觉内容的开始和结束。为了能够生成文本和图像，LaVIT 采用两种图文连接形式：[image, text] 和 [text; image]。

对于这些多模态输入序列，LaVIT 采用统一的、自回归方式来直接最大化每个多模态序列的似然性进行预训练。这样在表示空间和训练方式上的完全统一，有助于 LLM 更好地学习多模态交互和对齐。在预训练完成后，LaVIT 具有感知图像的能力，可以像处理文本一样理解和生成图像。

3．实验结果

零样本多模态理解：如表 1 所示，LaVIT 在图像字幕生成（NoCaps、Flickr30k）和视觉问答（VQAv2、OKVQA、GQA、VizWiz）等零样本多模态理解任务上取得了领先的性能。

零样本多模态生成：在这个实验中，由于所提出的视觉 tokenizer 能够将图像表示为离散化 token，LaVIT 具有通过自回归生成类似文本的视觉 token 来合成图像的能力。作者对模型进行了零样本文本条件下的图像合成性能的定量评估，比较结果如表 2 所示，从表中可以看出，LaVIT 的表现优于所有其他多模态语言模型。与 Emu 相比，LaVIT 在更小的 LLM 模型上取得了进一步改进，展现了出色的视觉-语言对齐能力。此外，LaVIT 在使用更少的训练数据的情况下，实现了与最先进的文本到图像专家 Parti 可比的性能。

Method	Image Captioning		Visual Question Answering			
	Nocaps	Flickr	VQAv2	OKVQA	GQA	VizWiz
Flamingo-3B (Alayrac et al., 2022)	-	60.6	49.2	41.2	-	28.9
Flamingo-9B (Alayrac et al., 2022)	-	61.5	51.8	44.7	-	28.8
OpenFlamingo-9B (Awadalla et al., 2023)	-	59.5	52.7	37.8	-	27.5
MetaLM (Hao et al., 2022)	-	43.4	41.1	11.4	-	-
Kosmos-1 (Huang et al., 2023)	-	67.1	51.0	-	-	29.2
Kosmos-2 (Peng et al., 2023)	-	80.5	51.1	-	-	-
BLIP-2 (Vicuna-7B) (Li et al., 2023)	107.5	74.9	-	-	41.3	25.3
BLIP-2 (Vicuna-13B) (Li et al., 2023)	103.9	71.6	-	-	32.3	19.6
CM3Leon-7B (Yu et al., 2023)	-	-	47.6	-	-	37.6
Emu (LLaMA-13B) (Sun et al., 2023)	-	-	52.0	38.2	-	34.2
Ours (LLaMA-7B)	**114.2**	**83.0**	**66.0**	**54.6**	**46.8**	**38.5**

表 1　LaVIT 模型的零样本多模态理解任务评估

Method	Model Type	FID(↓)
Text2Image Specialist:		
DALL-E (Ramesh et al., 2021)	Autoregressive	28.0
CogView (Ding et al., 2021)	Autoregressive	27.1
SD (Rombach et al., 2022)	Diffusion	12.6
GLIDE (Nichol et al., 2021)	Diffusion	12.2
DALL-E2 (Ramesh et al., 2022)	Diffusion	10.4
Make-A-Scene (Gafni et al., 2022)	Autoregressive	11.8
MUSE-7.6B (Chang et al., 2023)	Non-Autoregressive	7.9
Imagen-3.4B (Saharia et al., 2022)	Diffusion	7.3
Parti-20B (Yu et al., 2022b)	Autoregressive	**7.2**
Multimodal Large Langauge Model:		
GILL (OPT-6.7B) (Koh et al., 2023)	LLM	12.2
Emu (LLaMA-13B) (Sun et al., 2023)	LLM	11.7
CM3Leon-7B (Yu et al., 2023)	LLM	10.8
Ours (LLaMA-7B)	LLM	**7.4**

表 2　LaVIT 模型的零样本文本到图像生成性能

图 7　LaVIT 模型的多模态图像生成结果示例

多模态提示图像生成：LaVIT 能够在无需进行任何微调的情况下，无缝地接受多种模态组合作为提示，生成相应的图像，而无需进行任何微调，如图 7。LaVIT 生成的图像能够准确反映给定多模态提示的风格和语义。而且它可以通过输入的多模态提示修改原始输入图像。在没有额外微调的下游数据的情况下，传统的图像生成模型如 Stable Diffusion 无法达到这种能力。

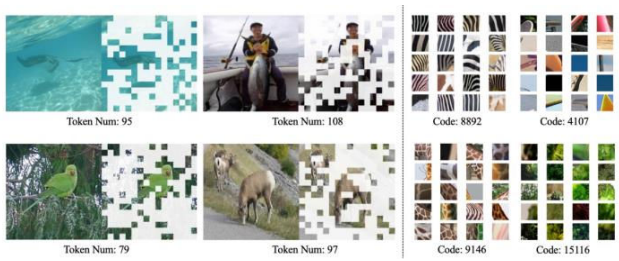


图 8　LaVIT 模型动态视觉分词器（左）和学习到的 codebook（右）的可视化示例

定性分析：如图 8 所示，LaVIT 的动态分词器可以根据图像内容动态选择最具信息量的图像块，学习到的代码本可以产生具有高层语义的视觉编码。

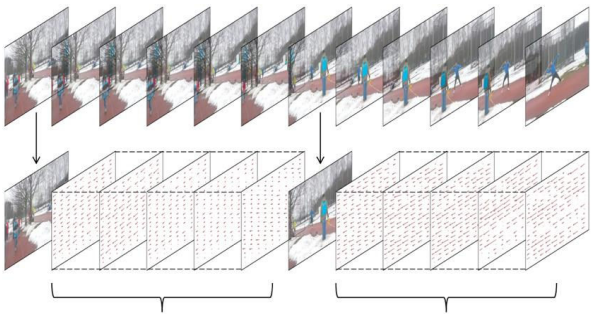


图 9　基于视觉运动解耦的视频表示

三、自回归文生视频大模型 Video-LaVIT

1．背景介绍

最近，大语言模型（LLMs）的重大突破引发了研究者们开发多模态大语言模型的热潮，已经出现了像GPT-4V，Gemini这样的多模态智能体，可以准确地理解图像中的内容。尽管取得了一定的成功，这些多模态大语言模型仍主要集中在图像-文本数据上，对于视频模态的探索则相对较少。与静态图像相比，视频作为一种动态的媒体形式，其更符合人类的视觉感知。因此，从视频数据中学习对于帮助智能体理解现实世界尤为重要。

视频理解的关键挑战在于：如何有效地对时空动态信息进行建模，例如随着时间变化的动作和场景等。目前，已经有一些方法尝试去利用语言模型(LLM)的强大推理能力来处理视频数据，它们将不同的视频帧当作不同的图像分别进行独立编码。然而，这种编码方式无法很好地捕捉时序信息。尽管最近的研究VideoPoet[21]尝试通过3D视频编码器来处理视频生成，但其适用性受限于短视频片段，因为其产生的长token序列（例如，VideoPoet对于一个2.2秒的视频片段需要使用1280个token进行编码）会导致计算资源的巨大消耗。

那么，如何以更加高效的方式在语言模型中编码视频呢？可以观察到，同一个视频镜头中的不同视频帧之间通常存在较多的时间冗余，没有必要将所有的帧都编码为输入到语言模型的token。因此，本文旨在寻找一种更高效的方法来编码视频中的时间运动信息，无需一次编码所有帧。如图9所示，我们将一个视频片段分解为交替的关键帧和运动向量，并在语言模型中进行分别编码。关键帧表示主要的视觉语义，而运动向量则表示基于关键帧的时间演变。这种解耦的视觉运动表示具有两个主要优势。首先，我们不需要编码所有视频帧来建模时间信息，只需编码相邻帧之间的差异。由于运动向量比密集的像素值更为稀疏，因此编码运动向量所需的token数量可以大大减少，这可以提高大规模视频预训练的效率。此外，解耦表示使得我们能够分别建模视觉和运动信息。关键帧就类似一张图像，所以我们可以自然地继承基于图像的模型的视觉知识，而不需要从头开始训练。

基于这种视觉运动解耦的表示，本文提出的Video-LaVIT模型将得到的关键帧和运动向量都分词为离散化的token，并以自回归的方式预测下一个图像、运动、或文本token，因此实现对视频、图像和文本的统一生成预训练。在训练完成后，Video-LaVIT具有对图像和视频内容的理解和生成能力。

2．模型架构

视频编码：Video-LaVIT 模型的核心在于将视频分解为关键帧和运动向量，关键帧捕捉主要的视觉语义，而运动向量描述其对应的关键帧随时间的动态演变。具体来说，在像 MPEG-4 这样的视频编码协议中，主要有两种帧类型：I 帧和 P 帧。I 帧是关键帧传达主要的视觉内容，并作为后续 P 帧的参考。P 帧仅保留与其前一帧

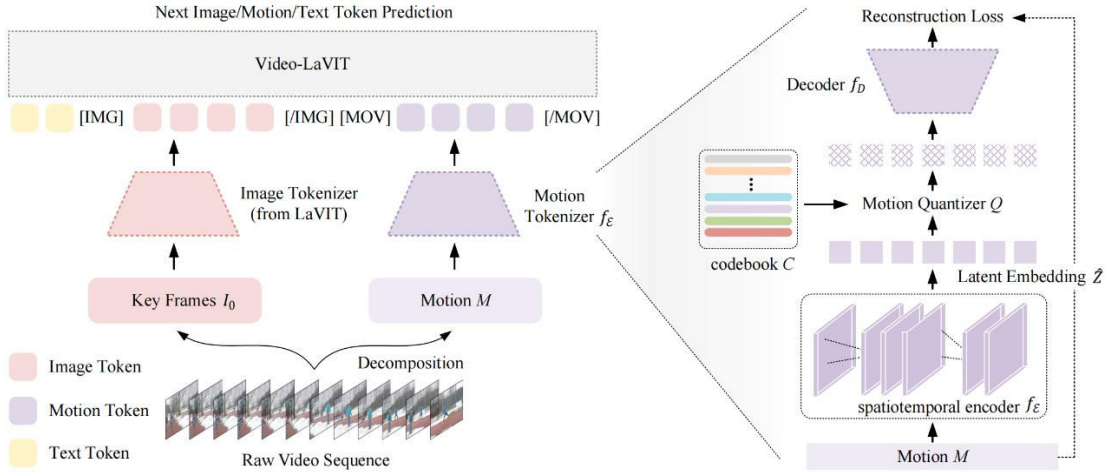


图 10　Video-LaVIT 模型的整体架构

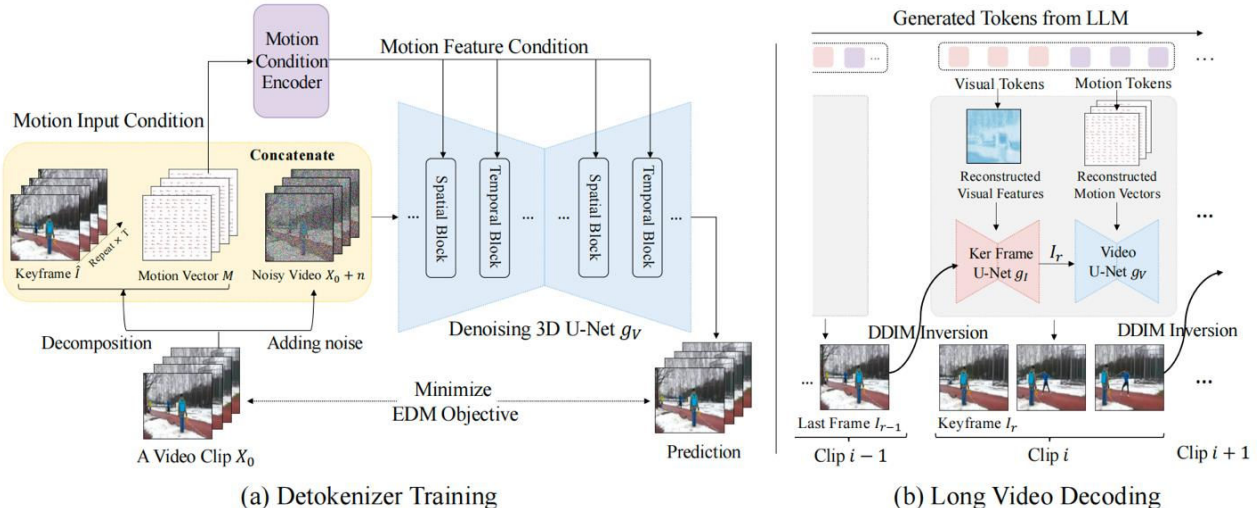


图 11　Video-LaVIT 模型的视频解码

的差异，而这些差异通过运动向量进行编码。具体来说，每个 P 帧被划分为 16×16 的不重叠宏块，其运动向量则是通过找到与相邻帧之间最佳块对应关系来估计的。简而言之，运动向量是视频中时序动作的高效表示。与每个像素位置的密集光流相比，它计算的是块之间的运动，并且可以直接从视频编解码器中高速提取。如图 10 所示，Video-LaVIT 将得到的关键帧和运动向量都分词为离散化的 token，和文本一起放在语言模型中进行统一建模。

视频解码：相对于上一节的编码，视频解码则是一个逆过程：将由多模态语言模型生成的离散化token映射回其原始的连续像素空间。考虑到学习从离散标记到高维视频空间直接映射的挑战，我们采用了顺序解码策略。如图11所示，首先解码关键帧，随后的其他帧则通过视频解码器恢复。视频解码器采用去噪U-Net架构，它通过以关键帧和运动向量作为条件，重建原始视频片段的所有视频帧。为了确保重建的视频帧严格遵循原本视频的运动信息，我们还设计了一种增强的运动编码策略。除了将运动向量作为解码器的直接输入外，我们还在解码器的block中增加了空间和时间交叉注意力层，加强对运动信号的进一步编码。整个视频解码器的训练则只需视频数据，不需要任何文本描述，因此可以扩展到更多无监督的视频数据。

由于视频被表示为多个交替的(visual, motion)序列，Video-LaVIT可以自然地通过自回归的方式生成多个视频片段，支持创建更长的视频。值得注意的一点是，如果分别对不同的视频片段进行解码，不同的视频片段之间会出现一些细粒度的视觉细节不一致的情况。为了解决这个问题，我们在解码视频片段时加入了一个显式的噪声约束。如图11所示，我们将最后一个片段的结束帧反转为中间噪声状态，然后将这个噪声状态作为下一个视频片段解码过程中的初始噪声。这样，相邻的视频片段就可以明确地相互关联起来，这对于长视频的生成至关重要。

多模态内容的联合预训练：基于本文提出的视觉-运动解构的离散化分词策略，我们可以不加区分地将所有模态（视频、图像和文本）视为输入到语言模型中的一系列离散token。这使得模型能够继承大语言模型成功的训练范式，以自回归的方式直接最大化每个标记的似然性。经过预训练后，Video-LaVIT能够生成不同模态的token，实现多模态的理解和生成。

3．实验结果

图像和视频理解：在11个常用的图像和视频基准测试中，Video-LaVIT展示了其在多模态理解上的能力。在图像理解上，其在八个广泛使用的图像问答和多模态基准测试中都取得了最佳的性能。例如，在SQA上，它比具有更高输入分辨率的LLaVA-1.5高出3.2%。在三个常见的视频基准测试中，Video-LaVIT与多个最近的视频-语言模型进行了比较，在这三个基准测试中均取得了最先进的性能。例如，在MSVD-QA上超过了之前领

Method	LLM size	Image Question Answering				Multimodal			
		VQA^{v2}	GQA	VizWiz	SQA^{I}	MME	MMB	SEED	MM-Vet
Flamingo (Alayrac et al., 2022)	9B	51.8	-	28.8	-	-	-	-	-
BLIP-2 (Li et al., 2023b)	13B	41.0	41.0	19.6	61.0	1293.8	-	46.4	22.4
InstructBLIP (Dai et al., 2023)	13B	-	49.5	34.3	63.1	1212.8	44.0	-	25.6
CM3Leon (Yu et al., 2023a)	7B	47.6	-	37.6	-	-	-	-	-
Emu (Sun et al., 2024)	13B	52.0	-	34.2	-	-	-	-	36.3
DreamLLM (Dong et al., 2024)	7B	72.9*	-	49.3	-	-	58.2	-	**36.6**
Video-LLaVA (Lin et al., 2023)	7B	74.7*	60.3*	48.1	66.4	-	60.9	-	32.0
LLaMA-VID (Li et al., 2023f)	7B	78.3*	63.0*	52.5	67.7	1405.6	65.3	59.7	-
LLaVA-1.5 (Liu et al., 2023a)	7B	78.5*	62.0*	50.0	66.8	1510.7	64.3	58.6	30.5
Video-LaVIT	7B	**80.3***	**64.4***	**56.0**	**70.0**	**1551.8**	**67.3**	**64.0**	33.2

Method	LLM size	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM (Yang et al., 2022)	1B	32.2	-	16.8	-	24.7	-
Video-LLaMA (Zhang et al., 2023)	7B	51.6	2.5	29.6	1.8	12.4	1.1
VideoChat (Li et al., 2023d)	7B	56.3	2.8	45.0	2.5	26.5	2.2
Video-ChatGPT (Maaz et al., 2023)	7B	64.9	3.3	49.3	2.8	35.2	2.7
LLaMA-VID (Li et al., 2023f)	7B	69.7	3.7	57.7	3.2	47.4	3.3
Video-LLaVA (Lin et al., 2023)	7B	70.7	3.9	59.2	**3.5**	45.3	3.3

表 3　Video-LaVIT 模型的图像和视频理解性能

先的模型Video-LLaVA 2.5%

文本和图像生成视频：通过在大规模图像视频数据上进行预训练，Video-LaVIT能够根据人类指令，以自回归的形式生成多个不同的视频片段。如图12所示，与商用的生成模型Gen-2相比，Video-LaVIT能够生成更复杂的物体运动，同时不违反物理规则的视频内容。

除了文本提示，Video-LaVIT还支持图像到视频的生成。给定一张图像输入，其可以将图像处理为离散的视觉标记，并输入到多模态语言模型中生成运动token。生成的运动信息与输入图像结合后，可以解码成一个视频片段。Video-LaVIT可以通过生成不同的合理运动来为输入图像添加动画效果。

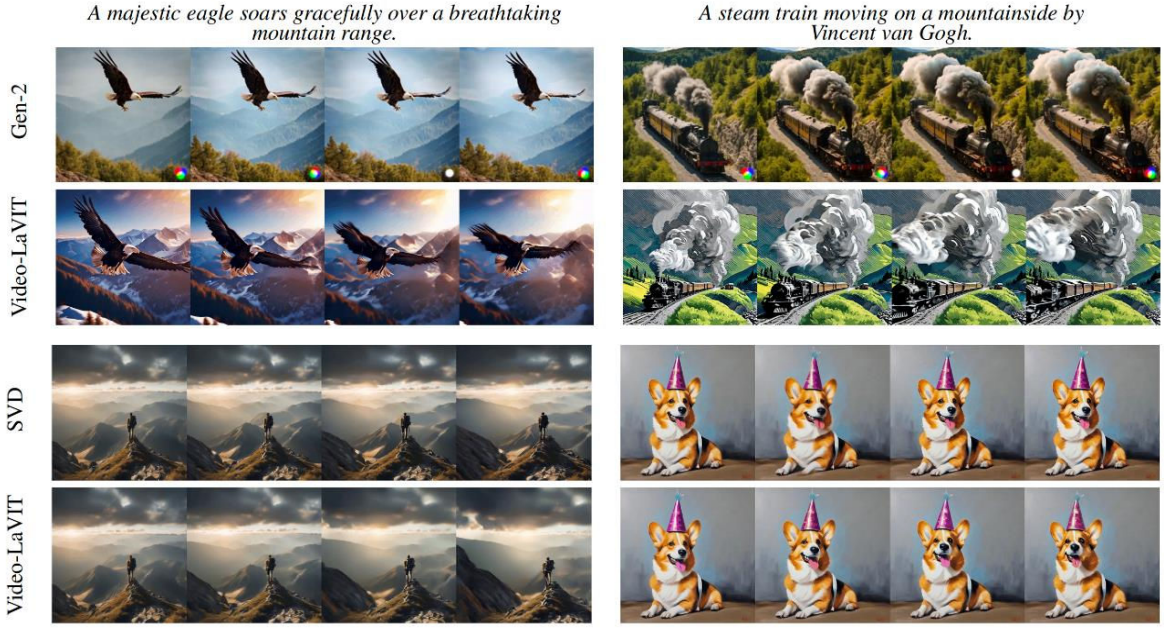


图 12　Video-LaVIT 模型根据文本或图像指令生成视频，与 Gen-2 和 SVD 的对比

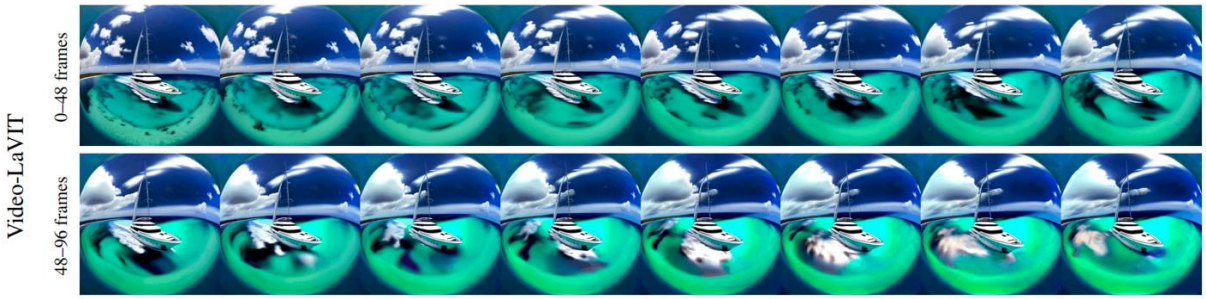


图 13　Video-LaVIT 模型的长视频生成效果

长视频生成：通过在解码连续视频片段时显式地约束噪声，Video-LaVIT能够在长视频生成中保持高度的时间一致性。如图13所示，所生成的视频片段之间的运动和视觉信息都具有高度的一致性。

四、总结

该论文提出了基于自回归架构的多模态生成式大模型 LaVIT 与 Video-LaVIT。其中 LaVIT 的提出为多模态任务的处理又提供了一种创新范式，通过使用动态视觉分词器将视觉和语言表示为统一的离散 token 表示，继承了 LLM 成功的自回归生成学习范式。通过在统一生成目标下进行优化，LaVIT 可以将图像视为一种外语，像文本一样理解和生成它们。Video-LaVIT 通过引入视觉与运动解耦，将该思路推广至文生视频任务。上述方法的成功为未来多模态研究的发展方向提供了新的启示，利用 LLM 强大的推理能力，实现更智能、更全面的多模态理解，并为生成打开新的可能性。

最后，本文介绍的 LaVIT、Video-LaVIT 的代码和模型均发布于：https://github.com/jy0205/LaVIT。

责任编委　**崔海楠**

参考文献

[1] Jin, Yang, et al. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In ICLR 2024.
[2] Jin, Yang, et al. Video-LaVIT: Unified video-language pre-training with decoupled visual-motional tokenization. In ICML 2024.
[3] Ramesh, Aditya, et al. Zero-shot text-to-image generation. In ICML 2021.
[4] Ramesh, Aditya, et al. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125, 2022.
[5] Betker, James, et al. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3, 2023.
[6] Ding, Ming, et al. CogView: Mastering text-to-image generation via transformers. In NeurIPS 2021.
[7] Rombach, Robin, et al. High-resolution image synthesis with latent diffusion models. In CVPR 2022.
[8] Nichol, Alex, et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In ICML 2022.
[9] Chang, Huiwen, et al. MaskGIT: Masked generative image transformer. In CVPR 2022.
[10] Ding, Ming, et al. CogView2: Faster and better text-to-image generation via hierarchical transformers. In NeurIPS 2022.
[11] Saharia, Chitwan, et al. Photorealistic text-to-image diffusion models with deep language understanding. In NeurIPS 2022.
[12] Yu, Jiahui, et al. Scaling autoregressive models for content-rich text-to-image generation. In TMLR 2022.
[13] Gafni, Oran, et al. Make-A-Scene: Scene-based text-to-image generation with human priors. In ECCV 2022.
[14] Ruiz, Nataniel, et al. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In CVPR 2023.
[15] Chang, Huiwen, et al. Muse: Text-to-image generation via masked generative transformers. In ICML 2023.

[16] Zhang, Lvmin, et al. Adding conditional control to text-to-image diffusion models. In ICCV 2023.
[17] Blattmann, Andreas, et al. Align your latents: High-resolution video synthesis with latent diffusion models. In CVPR 2023.
[18] Blattmann, Andreas, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127. 2023.
[19] Wu, Chenfei, et al. NÜWA: Visual synthesis pre-training for neural visual world creation. In ECCV 2022.
[20] Girdhar, Rohit, et al. Emu video: Factorizing text-to-video generation by explicit image conditioning. In ECCV 2024.
[21] Kondratyuk, Dan, et al. VideoPoet: A large language model for zero-shot video generation. In ICML 2024

金 阳

金阳，北京大学前沿交叉学科研究院 2022 级博士研究生，导师为穆亚东研究员，主要研究方向为多模态大语言模型,视频理解。
Email: jiny@stu.pku.edu.cn

孙至诚

孙至诚，北京大学前沿交叉学科研究院 2020 级博士研究生，导师为穆亚东研究员，主要研究方向为持续学习，多模态生成。
Email: sunzc@pku.edu.cn

穆亚东

穆亚东，北京大学研究员、长聘副教授、博士生导师、北大博雅青年学者，先后在北京大学获得理学学士和理学博士学位。曾在新加坡国立大学、美国哥伦比亚大学、华为香港诺亚方舟实验室、美国电话电报公司研究院（AT&T Labs）担任研究职位，入选国家级人才计划，在国际主流会议和期刊发表论文 120 余篇，其中 CCF 推荐 A 类会议和 ACM/IEEE 汇刊论文 80 余篇，申请国内外专利 30 余项。获得陕西省自然科学一等奖和国际会议 SIGIR 最佳论文提名奖。担任多媒体领域旗舰期刊 IEEE Transactions on Multimedia 的编委，多次担任计算机视觉领域顶级会议（如 CVPR、ACM Multimedia）的领域主席。
Email: myd@pku.edu.cn